DOCUMENT RESUME

ED 223 662 TM 820 786

AUTHOR	Sympson, James B.; And Others
TITLE	Predictive Validity of Conventional and Adaptive Tests in an Air Force Training Environment. Interim Report.
INSTITUTION	Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO	AFHRL-TR-81-40
PUB DATE	Aug 82
CONTRACT	F33615-77-C-0061
NOTE	120p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Ability Identification; *Aptitude Tests; *Computer Assisted Testing; Generalization; Latent Trait Theory; *Personnel Selection; *Predictive Validity; Testing; Test Validity
IDENTIFIERS	*Armed Services Vocational Aptitude Battery; Criterion Related Validity; *Tailored Testing; Three Parameter Model

ABSTRACT

Conventional Armed Services Vocational Aptitude Battery-7 (ASVAB) Arithmetic Reasoning and Word Knowledge tests, were compared with computer-administered adaptive tests as predictors of performance in an Air Force Jet Engine Mechanic training course (n=495). Results supported earlier research in showing somewhat longer examinee response times for adaptive tests in comparison to conventional tests. These longer response times were attributed to the higher relative difficulty of the items in the adaptive tests. Score information analyses showed that the adaptive tests provided considerably higher levels of information than did the conventional tests at all ability levels. Analyses of composite validities also showed on significant effects involving adaptive versus conventional tests, although there was again a siginificant interaction involving the adaptive tests. The data thus indicated no significant differences in validities between equal-length adaptive and conventional tests. It is concluded that similar validities for adaptive tests and conventional tests are supportive of the use of adaptive tests in military selection testing because of additional advantages inherent in computerized adaptive administration of ability tests. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Air Force Human Resources Laboratory

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

AFHRL-TR-81-40

AIR FORCE HUMAN RESOURCES LABORATORY

ED223662

TM 820 986

PREDICTIVE VALIDITY OF CONVENTIONAL AND ADAPTIVE TESTS IN AN AIR FORCE TRAINING ENVIRONMENT

By

James B. Sympson
David J. Weiss

Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

Malcolm James Ree

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

August 1982

Interim Report for Period July 1977 — July 1981

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

MALCOLM JAMES REE
Contract Monitor

NANCY GUINN, Technical Director
Manpower and Personnel Division

RONALD W. TERRY, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TR-81-40

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
PREDICTIVE VALIDITY OF CONVENTIONAL AND ADAPTIVE TESTS IN AN AIR FORCE TRAINING ENVIRONMENT

5. TYPE OF REPORT & PERIOD COVERED
Interim
July 1977 — July 1981

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
James B. Sympson
David J. Weiss
Malcolm James Ree

8. CONTRACT OR GRANT NUMBER(s)
F33615-77-C-0061

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61101F 62703F
ILIR0013 77191805

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
August 1982

13. NUMBER OF PAGES
120

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

ability testing
adaptive testing
Bayesian scoring
Bayesian test
criterion-related validity
item response theory
latent trait theory
linear-model analysis
logistic test model
maximum information test
maximum likelihood scoring
stratified maximum information test
tailored testing
three-parameter model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Conventional ASVAB-7 Arithmetic Reasoning (AR) and Word Knowledge (WK) tests were compared with computer-administered adaptive tests as predictors of performance in an Air Force Jet Engine Mechanic (JEM) training course. All test items were calibrated using a 3-parameter logistic item response model. Adaptive tests were composed of items selected by Bayesian and stratified maximum information (STMI) strategies. Each of 495 JEM trainees was administered three AR tests and three WK tests by computer, including two adaptive tests (Bayesian and STMI) and one conventional ASVAB test in each ability area. Since the two adaptive tests selected items from the same item pools, in each ability area a special item fill-in procedure was used during an examinee's second adaptive test to eliminate repeated presentation of the same items. The conventional ASVAB tests were scored by number-correct.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Item 20 (Continued):

Bayesian, and maximum likelihood methods. Order of test administration was counterbalanced in subgroups of examinees. Validity data were analyzed by multivariate linear-model analyses with testing-strategy/scoring-method (TSSM), item type (AR, WK) and/or order of administration as independent variables, and single-test and composite validities as dependent variables. Additional dependent variables studied included characteristics of ability estimates (distributions, intercorrelations, and score information functions), examinee and computer response times, and the effect of fixed versus variable entry.

Results supported earlier research in showing somewhat longer examinee response times for adaptive tests in comparison to conventional tests. These longer response times were attributed to the higher relative difficulty of the items in the adaptive tests. Score information analyses showed that the adaptive tests provided considerably higher levels of information than did the conventional tests at all ability levels.

The linear-model analysis of single-test validities showed a significant three-way interaction among TSSM, item type, and order of administration, as well as main effects for item type and TSSM. Analysis of the main-effect contrasts for TSSM showed that maximum likelihood scoring of ASVAB resulted in lower validities than either the number-correct or Bayesian scoring of ASVAB. None of the contrasts involving adaptive versus conventional tests were significant. Analysis of the three-way interaction contrasts showed that this interaction was due to differences between the adaptive strategies.

Analyses of composite validities also showed no significant effects involving adaptive versus conventional tests, although there was again a significant interaction involving the adaptive tests. The data thus indicated no significant differences in validities between equal-length adaptive and conventional tests. The data did indicate, however, that the STMI adaptive testing strategy could achieve validities that approximated those of the ASVAB tests while requiring only one-third to one-half the number of items. It is concluded that similar validities for adaptive tests and conventional tests are supportive of the use of adaptive tests in military selection testing because of additional advantages inherent in computerized adaptive administration of ability tests.

SUMMARY

Objective

To compare the predictive validity of two computerized adaptive testing (CAT) strategies with the predictive validity of conventional Armed Services Vocational Aptitude Battery (ASVAB) subtests in a military training environment.

Background

In the last 10 years, CAT has emerged as a means of improving the quality of ability measurements for armed services personnel. Previous research on CAT has been primarily concerned with investigations of the accuracy and precision of ability estimates derived from various methods of implementing CAT in comparison to conventional ability tests. The few validity studies comparing CAT and conventional tests have not examined the predictive validity of CAT in a military training environment. In addition, previous research has been restricted to a single ability domain and has not directly compared the validities of different adaptive testing strategies. The present study was designed to investigate the validity of two CAT strategies using two ability domains.

Approach

Conventional and adaptive tests were administered to a large group of Air Force recruits who were beginning a Jet Engine Mechanic (JEM) training course. The validity criterion was their final grade at the end of the course.

Specifics

Method. Each of 495 JEM trainees was administered three Arithmetic Reasoning (AR) tests and three Word Knowledge (WK) tests by computer. Bayesian and stratified maximum information (STMI) adaptive tests and conventional ASVAB subtests were administered in each ability domain. To control warmup and fatigue effects, the order of test administration was counterbalanced in subgroups of the examinees.

In each ability domain, adaptive tests selected items from the same item pool. A special item fill-in procedure was used so that duplicate items would not be administered in the adaptive tests. All test items were calibrated, and the adaptive testing strategies were implemented, using Birnbaum's three-parameter logistic item response model. In addition to conventional number-correct scores, Bayesian and maximum likelihood ability estimates were also generated for the ASVAB subtests.

Validity data were analyzed by multivariate linear-model analyses with testing-strategy/scoring-method (TSSM), item type (AR, WK), and/or order of administration as independent variables and with single-test and composite validities as dependent variables. Other dependent variables included the distributional and information characteristics of the ability estimates from the various TSSM combinations, computer and examinee response times, and the effect of fixed versus variable entry in the adaptive tests.

Findings and discussion. The results showed longer examinee response times for the adaptive tests in comparison to the conventional tests. This result was expected because the adaptive test presents items which are on the average more difficult for the examinee thereby prohibiting exclusion of the easy and the difficult items which may be answered quickly. However, in field application, the adaptive test would employ fewer items than the conventional test so that testing time should not exceed that of conventional tests. The adaptive tests were of more appropriate item difficulty for the examinees than were the conventional tests. Information analyses showed that the adaptive tests provided considerably higher levels of test score information at all ability levels than did the conventional tests.

Results of the validity analyses, both at the single-test level and for test-score composites, showed no significant differences in validities between the conventional and adaptive tests. There were some differences between the validities of the adaptive tests as they interacted with order of test administration and item type. The maximum likelihood scoring of ASVAB resulted in significantly lower validities than did Bayesian or number-correct scoring.

Although there were no significant differences in validities between full-length adaptive and conventional tests, under the fixed-entry condition the STMI strategy achieved validities that approximated those of the ASVAB tests while requiring only one-third to one-half the number of items.

Conclusions

The fact that mean levels of criterion-related validity were not significantly different for the adaptive and ASVAB tests is interpreted as supportive of the use of CAT in military testing, due to the other advantages inherent in CAT. These advantages include higher levels of measurement precision, immediate scoring of tests, immediate availability of test scores for use in making military assignment decisions, and alleviation of the test compromise problem.

PREFACE

This study was conducted under task 771918, Selection and Classification Technologies. The research focuses on the development of procedures and techniques to refine and improve measurement devices used in the Air Force operational testing program.

This work is part of a continuing series of studies to evaluate the efficacy of computer driven adaptive testing. The effort supports the subthrust area of Assessment of Personnel Qualifications, under the major thrust area of Manpower and Force Management.

CONTENTS

Introduction 1
  Purpose 2

Method 2
  Calibration of Test Items 2
    Arithmetic Reasoning 3
    Word Knowledge 6
      Expanding the WK Item Pool 7
    ASVAB Calibrations 10
      Arithmetic Reasoning 10
      Word Knowledge 12
  Testing Strategies and Scoring Methods 12
    Adaptive Tests 12
      Bayesian 12
      Stratified Maximum Information 14
    ASVAB 16
      Bayesian Scoring 16
      Maximum Likelihood Scoring 16
      Number-Correct Scoring 17
  Adaptive Test Entry Procedures 17
    Fixed Entry 17
      Estimating the Ability Distributions 17
      Fixed-Entry Ability Estimates 19
    Variable Entry 21
  Apparatus 23
    Testing System Hardware 23
    Test Administration Software 24
    System Checkout and Installation 24
  Subjects 25
    The Experimental Sample 25
    Cases Retained for Analyses 25
  Data Collection Procedures 27
    The Testing Environment 27
    Instructions 27
    Test Administration 28
      Counterbalancing of Orders 29
      Item "Fill-in" Procedure 30
  The Criterion 31
  Design 33
    Independent Variables 33
      Analyses for Single Tests 33
      Analyses for Composites 34
    Dependent Variables 34
      Test Score 34
      Mean Examinee Response Time 34
      Mean Computer Response Time 35
      Single-Test Validity 36
      Composite-Test Validity 36

CONTENTS (continued)

Data Analysis Procedures 36
The Adaptive Test Fill-in Procedure 36
Characteristics of Ability Estimates 37
Out-of-Bounds and Non-Converged Maximum Likelihood Estimates 37
Distributions and Correlations 37
Information 38
Response Times 38
Mean Examinee Response Time 38
Mean Computer Response Time 38
Evaluation of Variable Entry Procedure 39
Validity Analyses 39
Single Tests 39
Sequential Validity Analysis 39
Multivariate Linear-Model Analysis 40
Cell Coding 42
Linear Contrasts 44
Estimating the Variance-Covariance Matrix 45
Test Statistics 47
A Posteriori Significance Tests 48
Validities for Composites 49
Fixed-Weight Composites 51
Optimally-Weighted Composites 52
Pre-Enlistment ASVAB Composites 54
Effect of Limiting Boundary Values on the Validity of Maximum Likelihood Ability Estimates 55
Intercorrelations Between AR and WK Tests 56

Results 56
Preliminary Results 56
Fill-In Analysis 56
Nonconverged and Out-of-Bounds Maximum Likelihood Estimates 56
Effect of Variable Entry 59
Examinee Response Time 60
Computer Response Time 60
Characteristics of Test Scores 66
Distributions 66
ASVAB Number Correct 66
IRT Ability Estimates 66
Correlations 71
Information 71
Validity 74
Single Tests 74
Validity as a Function of Test Length 74
Effect of Maximum Likelihood Boundary Values 74
Linear-Model Analysis 77

CONTENTS (continued)

Composites 82
Intercorrelations between AR and WK scores 82
Linear-Model Analyses 83
Comparison with Pre-Enlistment ASVAB Composites 85

Discussion and Conclusions 86
Score Information 88
Validity 89
Relationship with Previous Research 90
Conclusions 94

References 95

Appendix: Supplementary Tables 101

LIST OF TABLES

Table | Page

1 Calibration of Arithmetic Reasoning Items in the Air Force Population 4

2 Centile Points of the Distributions of Estimated Item Parameters in the AR and WK Item Pools 6

3 Item Parameter Estimates for Items in ASVAB-7 AR and WK Subtests 11

4 Experimental Subgroups Formed by Applying Alternative Retention Criteria 26

5 Order Conditions for Test Administration 29

6 Coded Vectors for Linear-Model Analysis of Single-Test Validities 44

7 Cell Codes in Two-Way Cross-Classification for Composite-Score Validities 50

8 Coded Vectors for Linear-Model Analyses of Composite-Score Validities 51

9 Relative Frequency in WK-AR and AR-WK Groups of Out-of-Bounds Low ($\hat{\theta} = -5.0$) and High ($\hat{\theta} = 5.0$) Maximum Likelihood Ability Estimates 60

10 Summary Statistics for Mean Examinee Response Time Distributions for ASVAB, BAYES, and STMI Tests with AR and WK Items, for WK-AR and AR-WK Groups 62

11 Summary Statistics for Mean Computer Response Time Distributions for ASVAB with AR and WK Items, for Adaptive-Test Order Groups 62

12 Summary Statistics for Mean Computer Response Time Distributions for BAYES and STMI Adaptive Tests with AR and WK Items 66

13 Summary Statistics for ASVAB/N, ASVAB/B, ASVAB/M, BAYES, and STMI Score Distributions with AR and WK Items, for WK-AR and AR-WK Groups 67

14 Pearson Product-Moment Correlations Among Test Scores, Separately for AR and WK Tests, in Two Subgroups 72

LIST OF TABLES (continued)

Table Page

15 Criterion-Related Validity Correlations of Maximum Likelihood Ability Estimates for Original-Bounds, In-Bounds, and Revised-Bounds Scoring 77

16 Criterion-Related Validity Correlations for Single Tests 78

17 Marginal TSSM Means for Single-Test Validities 79

18 Three-Way Linear-Model Analysis for Single-Test Validities . 79

19 Three-Way Interaction Contrasts for Single-Test Validities . 80

20 Pairwise Contrasts Among Marginal TSSM Means for Single-Test Validities 81

21 Pearson Product-Moment Correlations Among Test Scores in the WK-AR and AR-WK Graduate Groups, for AR and WK Items . 82

22 Criterion-Related Validity Correlations for Fixed-Weight and Optimally-Weighted Composites 84

23 Two-Way Linear-Model Analyses for Fixed-Weight-Composite and Optimally-Weighted-Composite Validities 84

24 Two-Way Interaction Contrasts for Fixed-Weight-Composite Validities 85

25 Criterion-Related Validity of Pre-Enlistment ASVAB Composite Scores and Experimental Fixed-Weight-Composite Scores 86

26 Alternate-Forms Reliability and Concurrent-Validity Correlations at a Test Length of 30 Items from Four Studies of Adaptive Testing 90

Appendix Table

A-1 Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive AR Tests in the Fixed-Entry Condition as a Function of Number of Items Administered, for AR-WK Group 101

LIST OF TABLES (continued)

Table Page

A-2 Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive WK Tests in the Fixed-Entry Condition, as a Function of number of Items Administered, for WK-AR Group 102

A-3 Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive AR Tests in the Variable-Entry Condition, as a Function of Number of Items Administered, for WK-AR Group 103

A-4 Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive WK Tests in the Variable-Entry Condition, as a Function of Number of Items Administered, for AR-WK Group 104

14

LIST OF FIGURES

Figure Page

1 Transformation of an Approximate ASVAB-7 Arithmetic Reasoning True-Score Distribution into a Theta Distribution via the Test Characteristic Curve 20

2 Cell Codes in the 3-Way Cross-Classification for Single-Test Validities 43

3 Percent of Examinees Having n Items Filled in During the Second Adaptive Test 57

4 Percent of Examinees with Nonconverged Ability Estimates at Each Stage in the STMI Tests 58

5 Number of Examinees with $\hat{\theta}$ of -5.0 at Each Stage of the STMI AR Test 59

6 Correlation of Interim Ability Estimates with Final Ability Estimates in the Variable-Entry and Fixed-Entry Conditions, for BAYES and STMI Tests 61

7 Frequency Distributions of Mean Examinee Response Time per Item in ASVAB, BAYES, and STMI AR Tests 63

8 Frequency Distributions of Mean Examinee Response Time per Item in ASVAB, BAYES, and STMI WK Tests 64

9 Frequency Distributions of Mean Computer Response Time per Item in ASVAB, BAYES, and STMI AR and WK Tests 65

10 Frequency Distributions of ASVAB AR and WK Number-Correct Scores for WK-AR and AR-WK Groups 68

11 Frequency Distributions of AR Ability Estimates for WK-AR and AR-WK Groups 69

12 Frequency Distributions of WK Ability Estimates for WK-AR and AR-WK Groups 70

13 Score Information Functions for Fixed-Entry STMI and BAYES Tests, an ASVAB-7 Test Adjusted to the Same Length as the Adaptive Tests, and Maximum Available Information 73

14 Score Information Functions for Three Scores on ASVAB-7 Subtests 75

LIST OF FIGURES (Continued)

Figure Page

15 Criterion-Related Validity Correlations of Fixed-Entry Adaptive Tests as a Function of Number of Items Administered, and Validity of ASVAB/N 76

16 Three-Way Interaction of TSSM, Item Type, and Order 80

17 Cross-Correlations Between Test Scores from AR and WK Tests 83

ACKNOWLEDGEMENTS

A number of individuals and organizations contributed to the successful completion of this project. Unfortunately, space limitations prevent mentioning them all here. Test data used in the item calibrations and in the item-parameter linking procedure were provided by the AFHRL Testing Detachment at Lackland AFB and the Technical Services Division of AFHRL. Development of the computerized testing system at the University of Minnesota depended heavily on the contributions of Douglas Larson, Warren Cartwright, and John Martin. Robert Gissing and Sandy Bowman of the USAF 3350th Technical Training Wing facilitated data collection activities at Chanute Air Force Base. Permission to use the SCAT items and a listing of 3-parameter logistic IRT parameters for these items were obtained from Frederic Lord at Educational Testing Service. Development of the linear-model procedures used in this research and preparation of most of this report were completed while the first author was at Educational Testing Service Inc., Princeton, NJ. The support of Frederic Lord and Albert Beaton, and the assistance of Al Rogers, John Jaskir and, especially, John Dunn during this period are greatly appreciated. The final phases of report preparation were completed while the first author was at the Navy Personnel Research and Development Center, San Diego, CA. To these individuals and organizations, and others not mentioned here, we extend our thanks.

PREDICTIVE VALIDITY OF CONVENTIONAL AND ADAPTIVE TESTS IN AN AIR FORCE TRAINING ENVIRONMENT

Previous research has compared computerized adaptive ability testing to conventional tests in a number of ways. Early research, comprised mainly of theoretical studies and simulation studies, demonstrated higher levels of measurement accuracy and precision for adaptive tests than for comparable conventional tests. These studies also showed that, even with substantial decreases in the length of adaptive tests, it was possible to obtain the same levels of measurement precision as with comparable conventional tests. Other studies, using live groups of examinees, reanalyzed data from conventional tests as though the data were from tests that had been adaptively administered. These studies supported the theoretical studies in showing potential decreases in test length, with no decreases in reliability, precision, or validity (see Weiss & Betz, 1973, for a review of pre-1973 research; also see, e.g., Vale, 1975, and McBride, 1979).

Early studies in which adaptive ability tests had been administered to real examinees were concerned primarily with the test-retest reliability of adaptive tests (e.g., Betz & Weiss, 1973, 1975; Larkin & Weiss, 1974; Vale & Weiss, 1975). These studies tended to show higher test-retest reliability for adaptive tests in comparison to conventional tests; however, the studies were limited to data obtained from college students.

More recently, live-testing validity studies have begun to appear. In one criterion-related validity study, Thompson and Weiss (1980) correlated scores on adaptive and conventional tests with grade-point averages (GPA) in groups of college students. Their data showed significantly higher correlations with GPA for some adaptive tests in comparison to conventional tests, even though the adaptive tests were substantially shorter than the comparable conventional tests.

In another live-testing study using college students, Kingsbury and Weiss (1980) correlated scores on alternate forms of conventional and adaptive tests with each other and with a 120-item conventional "criterion" test. Although their results showed higher alternate-forms reliabilities for the adaptive tests, the conventional test had higher correlations with the criterion test than did the adaptive test, at test lengths from 5 to 30 items.

McBride (1980) reported the first validity study of adaptive tests in comparison to conventional tests using a military recruit population. His data showed higher "concurrent validities" (correlations with a 50-item criterion test) for adaptive tests at test lengths up to 10 items, but equal or slightly higher validities for conventional tests from 15 to 30 items in length. A replication of McBride's study by Martin, McBride, and Weiss (in press), using a sample drawn from the same military recruit population, the same item pools and predictor tests, and the same 50-item criterion test, showed the concurrent-validity correlations for adaptive tests to be uniformly higher than for conventional tests for all test lengths from 1 to 30 items. Their data showed that an adaptive test of 11 items had a concurrent-validity correlation equivalent to a conventional test of 30 items.

While many of the results obtained in these previous validity studies are supportive of the use of adaptive tests, their generality with regard to practical applications of adaptive testing in a military environment is somewhat limited. First, the criteria used by Kingsbury and Weiss, by McBride, and by Martin et al. were long conventional tests containing the same types of items used in the adaptive and conventional predictor tests. Thus, the validity evidence presented was not predictive validity based on an operational military criterion. The studies reported by Kingsbury and Weiss and by Thompson and Weiss used college students, not military enlistees. Further, all four studies used only verbal ability items in the predictor tests, thus limiting their generality to that ability domain. Finally, with the exception of the Thompson and Weiss study, the studies were restricted to the use of only one adaptive testing strategy, and none of the four studies used an operational ASVAB subtest as the conventional predictor test. Thus, research on the criterion-related validity of adaptive tests in comparison with operational ASVAB subtests is needed.

Purpose

The present study was designed to investigate the criterion-related validity of adaptive tests using ASVAB types of items in comparison to operational ASVAB tests. Two ability domains were used--Word Knowledge (WK) and Arithmetic Reasoning (AR)--to study the generality of the findings across ability domains, and to investigate the validity of composites derived from combining the ability subtests in different ways. The criterion used was an operational military-training criterion obtained from a group of Air Force recruits who were tested in the early stages of their technical training. Finally, more than one adaptive testing strategy was used in order to investigate the generality of findings across different methods of adaptive testing.

METHOD

Calibration of Test Items

Test items were calibrated using Birnbaum's (1968, p. 405) three-parameter logistic model. This item response theory (IRT) model assumes that the probability of a correct response to an item is given by the function

$$P(\theta) = c + \frac{1 - c}{1 + e^{[-1.7a(\theta - b)]}} , \qquad [1]$$

where $\underline{a}$ is an item discrimination parameter,
$\underline{b}$ is an item difficulty parameter,
$\underline{c}$ is a lower asymptote parameter,
$\underline{\theta}$ indexes the person's level of ability,
and $\underline{e}$ is a constant (approximately 2.718).

Two item pools were calibrated for the experimental adaptive tests--an AR item pool and a WK item pool.

Arithmetic Reasoning

The AR item pool was comprised of quantitative reasoning problems of the type commonly found in tests of mathematical aptitude (e.g., Armed Forces Vocational Testing Group, 1974, p. 9; Educational Testing Service, 1978, pp. 12-14). A total of 264 four-alternative multiple-choice AR items were obtained from the Air Force Human Resources Laboratory (AFHRL); these items had been previously prepared and assembled into two 132-item test booklets designated PE7701 and PE7702.

Since all tests given in this research were to be administered by character-mode cathode-ray tube (CRT) terminals, six items in PE7701 and 10 items in PE7702 that required accompanying diagrams were not included in the calibration of the AR item pool. AFHRL provided item response data that had previously been obtained from Air Force enlistees who were administered either one-half or all of either PE7701 or PE7702 during their period of basic military training at Lackland AFB. Table 1 shows the nature of the data available for the AR item-pool calibration.

Table 1a shows that 842 Air Force enlistees had taken the first half of PE7701, 914 enlistees had taken the second half, and 155 enlistees had taken the entire test. Of 66 items appearing in the first half of the test, 61 were usable for this research. Of 66 items appearing in the second half of the test, 65 were usable. One examinee who answered all attempted questions correctly and three examinees who attempted fewer than 32 questions were eliminated from the PE7701 calibration. Because some of the remaining examinees omitted or did not reach various items, data were available from a minimum of 849 to a maximum of 1,040 individuals for the items calibrated in PE7701.

Table 1b gives the sample sizes and the number of items calibrated in PE7702. Sixteen examinees who attempted fewer than 32 questions were eliminated from this calibration. As for PE7701, some of the remaining examinees omitted or did not reach some of the items, resulting in calibration sample sizes of 819 to 939 individuals for the items in PE7702. Of the 248 items calibrated in PE7701 and PE7702, 209 were calibrated using the responses of at least 900 enlistees.

The computer program LOGIST (Wood, Wingersky, & Lord, 1976) was

Table 1
Calibration of Arithmetic Reasoning Items
in the Air Force Population

(a) Test PE7701

Examinees	Items K = 61	 K = 65
N = 842	PE7701 (1st half)	(not reached)
N = 914	(not reached)	PE7701 (2nd half)
N = 155	PE7701 (1st half)	PE7701 (2nd half)
Total = 1,911	N = 997	N = 1,069

(b) Test PE7702

Examinees	Items K = 62	 K = 60
N = 878	PE7702 (1st half)	(not reached)
N = 847	(not reached)	PE7702 (2nd half)
N = 129	PE7702 (1st half)	PE7702 (2nd half)
Total = 1,854	N = 1,007	N = 976

used to estimate item parameters for the AR data. Several computer simulation studies have demonstrated that LOGIST provides efficient estimates of the parameters of the three-parameter logistic model (Lord, 1975b; Ree, 1979; Swaminathan & Gifford, 1980; Sympson, 1977a).

LOGIST allows the user to code item responses as "correct," "incorrect," "omitted," and "not reached." Omitted responses (omits) are not merely treated as incorrect by LOGIST. Rather, omits influence a modified likelihood function that is maximized by the program in the process of estimating person abilities and item parameters. Lord (1974) has shown that use of this modified likelihood function gives ability estimates that converge, as test length increases, to the maximum likelihood estimates that would be obtained if omitted responses were replaced by random responses. Lord also showed that this method of estimation provides ability estimates that have smaller asymptotic variance about an examinee's true level of ability than would be obtained if omits were actually replaced by random responses.

LOGIST ignores responses coded as "not reached," no matter where they appear in the response vector. This allows construction of data files such as those represented in Table 1. For individuals taking only the second half of PE7701 or PE7702, the first 61 or 62 positions of their constructed response vectors were coded "not reached." In addition, items in the second half of the test that followed the last item attempted were also coded "not reached." For individuals taking only the first half of PE7701 or PE7702, all items after the last item attempted in the first half were coded "not reached."

The small group of individuals that took both halves of PE7701 served to link together the two larger groups that were exposed to only half of that test. In calibrating PE7701, LOGIST transformed the ability estimates ($\hat{\theta}$) for all the examinees so that the mean and standard deviation of $\hat{\theta}$ were approximately zero and one, respectively. This automatically defined the origin and unit of measurement for the $\hat{\theta}$ scale and, thus, the metric in which the item parameters $\underline{a}$ and $\underline{b}$ were expressed. The same process was followed in a separate computer run for calibrating the items in PE7702.

Since the available data did not include any individuals who took all or part of both PE7701 and PE7702, there was no way to be certain that the estimated IRT discrimination and difficulty parameters obtained in the two separate LOGIST runs were expressed in terms of a common metric. However, if it is assumed that the group of 1911 individuals who took all or part of PE7701 was a random sample from the same population as the group of 1854 individuals who took all or part of PE7702, then the metrics established in the two calibrations can be expected to differ by only small amounts arising through sampling error with respect to the mean and standard deviation of θ (true ability) in the two samples. For example, with a sample size of $N = 1900$ in each calibration group and a θ standard deviation of .95 in the population, the θ means in the two groups will differ by less than .051 in approximately 90% of such pairs of samples. Similarly, the standard deviations of θ will differ by less than .035 in approximately 90% of such pairs of samples. In view of these considerations, the item parameters obtained in the two separate AR calibrations were treated as if they had been expressed in terms of the same metric.

Following calibration of the 248 usable items in PE7701 and PE7702, two more items were eliminated from the pool. One of these was eliminated because its estimated discrimination parameter was below an arbitrarily selected cut-off value ($\underline{a} = .25$), while the other was eliminated when it was discovered that the item had been miskeyed in the original answer key for its test booklet. Thus, a total of 246 items were included in the AR item pool used in this research.

Table 2 provides information about the distributions of estimated item discrimination ($\underline{a}$), difficulty ($\underline{b}$), and lower asymptote ($\underline{c}$) parameters in the final AR item pool. Items with $\underline{a} < .25$ were not included. An upper limit of 2.00 was imposed on the estimated $\underline{a}$ parameters by LOGIST; eight items had discrimination parameters at this upper limit. Ninety percent of the item difficulty parameters fell between -1.533 and 1.876. A large number (166) of the $\underline{c}$ parameters were set to either .185 or .195 by LOGIST. These default values were calculated by LOGIST in the PE7701 and PE7702 calibration runs and assigned to items found to be too easy to allow accurate estimation of their lower asymptote parameters.

The Pearson product-moment correlation between the $\underline{a}$ and $\underline{b}$ parameters in the AR item pool was .493. This indication of a significant

Table 2
Centile Points of the Distributions of Estimated
Item Parameters in the AR and WK Item Pools

	AR			WK		
Centile	a	b	c	a	b	c
Minimum	.252	-4.159	.027	.251	-6.150	.024
5%	.445	-1.533	.120	.389	-3.663	.083
10%	.513	-1.274	.140	.472	-2.769	.110
20%	.633	-.626	.185	.570	-2.007	.130
30%	.723	-.220	.185	.673	-1.394	.150
40%	.802	.107	.185	.785	-.866	.150
50%	.929	.338	.195	.894	-.376	.155
60%	1.011	.631	.195	1.033	.136	.155
70%	1.118	.945	.195	1.132	.607	.155
80%	1.277	1.210	.195	1.250	1.640	.155
90%	1.491	1.502	.200	1.496	2.366	.195
95%	1.773	1.876	.233	1.837	2.984	.204
Maximum	2.000	2.594	.300	2.000	5.223	.316

positive relationship between the estimated item discrimination and difficulty parameters is consistent with results observed in other empirical item calibrations using LOGIST (Lord, 1975a). Lord (1975a) also reported the results of a LOGIST calibration using artificial data from simulated items. The items had a wide range of b values and had a values that were not systematically related to the level of item difficulty. The estimated a and b values obtained in the calibration were virtually unrelated. This suggests that the tendency for empirical a and b estimates to covary is not a result of some artifact in the LOGIST estimation procedure.

Word Knowledge

The WK item pool was composed of five-alternative multiple-choice WK items. Each item required the examinee to identify the one word among the five response alternatives which had a meaning that was most similar to the meaning of the item's stem word.

A total of 200 WK items were obtained from AFHRL. These items had previously been assembled into a single test booklet and administered to 1570 Air Force enlistees during their basic military training. Seventeen of these items were not included in the calibration of WK items for this research. These 17 items were dropped for a variety of reasons, including typographical errors present in an item at the time the calibration data were gathered, a negative item-total biserial correlation obtained in a previous AFHRL analysis of the calibration data, and the presence of two responses that could be defended as correct.

No examinee attempted fewer than 50 WK items, and no examinee answered correctly all the items he/she attempted; therefore data for all 1570 enlistees were retained for the WK item calibration. Because some examinees omitted or did not reach some items, sample sizes ranged from 1516 to 1535 individuals for each WK item calibrated. After LOGIST was used to calibrate the 183 usable AFHRL items, 15 of these were eliminated because their discrimination parameters were below .25. This left 168 items in the WK pool. Because a larger pool was desired, additional items were sought.

Expanding the WK item pool. In order to increase the size of the WK item pool, 120 five-alternative, multiple-choice WK items that had previously appeared in four different forms of the School and College Ability Tests (SCAT--Educational Testing Service, 1958) were utilized. Seven of these items were not used because their stem words duplicated stem words in the AFHRL WK item set. This left 113 usable SCAT items.

IRT parameters for the four SCAT forms had previously been estimated at Educational Testing Service (ETS) using data obtained from 1700, 2449, 2957, and 2998 individuals, respectively. Because some examinees omitted or did not reach various items, the actual number of responses available for calibrating each item in a given SCAT form was somewhat less than the number of available test records.

Since the LOGIST calibrations of the SCAT items at ETS had utilized samples of high-school students, the metric on which the SCAT a and b parameters were expressed was different from the metric of the AFHRL item parameters. To transform the ETS difficulty and discrimination parameters to the Air Force enlistee ability metric, a special test booklet was created and administered to a new sample of Air Force enlistees.

Sixty of the AFHRL WK items calibrated for this research were used as the odd-numbered items in the booklet. Sixty of the SCAT items previously calibrated at ETS were used as the even-numbered items in the booklet. Items selected for this "linking" test were chosen to span the entire range of available item difficulties. At each level of difficulty, the most discriminating AFHRL and ETS items were used.

Copies of the linking test were forwarded to AFHRL and the test was administered to 514 Air Force enlistees undergoing basic military training at Lackland AFB. Maximum likelihood ability estimates for each examinee were then generated using the odd and even halves of this test. At this point, it was discovered that two of the items included in the booklet were among the seven SCAT items that duplicated AFHRL item stems; consequently, the two maximum likelihood ability estimates computed for each person were based on 58 odd-numbered items and 58 even-numbered items.

After eliminating two individuals whose ability estimates did not converge within the interval between -5.0 and +5.0, and one "outlier" with an ability estimate of $\hat{\theta} \simeq -4.9$ on the SCAT ability scale, summary statistics were computed for the two sets of 511 ability estimates. The mean $\hat{\theta}$ from the odd-numbered (AFHRL) items was -.022. The standard deviation of these $\hat{\theta}$ values was 1.027. These values suggest that the new sample of enlistees could be viewed as coming from the same population as the group of 1570 enlistees used for calibrating the AFHRL WK items. The mean and standard deviation of $\hat{\theta}$ from the even-numbered (SCAT) items were .478 and .710, respectively. The difference (.500) between the SCAT mean and the mean ability estimate obtained using the AFHRL items indicated that the average level of ability in the original SCAT high-school samples would be about .5 unit (on the Air Force enlistee ability scale) below the mean ability of Air Force enlistees. Dividing the standard deviation obtained using the AFHRL items (1.027) by the standard deviation of the $\hat{\theta}$ from the SCAT items (.710) indicated that the standard deviation of each SCAT calibration sample would be about 1.45 on the Air Force enlistee ability scale. Thus, as a group, the Air Force enlistee population appeared to have a higher average level of ability, but a more restricted range, than did the SCAT high-school calibration samples.

When the assumptions of IRT are met, ability estimates obtained from two different sets of items calibrated in two different populations will differ by a linear transformation plus a component due to errors in estimating θ with item sets of finite length (Lord & Novick, 1968, pp. 379-381). Evidence for the applicability of IRT to the AFHRL and SCAT WK items used in this research was found in the bivariate scatterplot of the 511 "odd" and "even" ability estimates. The observed regression of either ability estimate on the other was clearly linear, and the Pearson product-moment correlation between the two sets of ability estimates was .889.

If a person's true ability on the Air Force enlistee ability metric is designated as θ_A and that same ability expressed on the SCAT calibration sample metric as θ_S, the relationship between θ_A and θ_S is

$$\theta_S = k_1\theta_A + k_2 \quad , \qquad [2]$$

where k_1 and k_2 are parameters of the linear transformation. Given this relationship, the expectation of θ_S can be expressed as

$$E(\theta_S) = k_1 E(\theta_A) + k_2 \ . \qquad [3]$$

Similarly, the variance of θ_S is

$$V(\theta_S) = (k_1)^2 \ V(\theta_A) \ . \qquad [4]$$

Thus

$$k_1 = \sqrt{V(\theta_S)/V(\theta_A)} \quad , \qquad [5]$$

and

$$k_2 = E(\theta_S) - k_1 E(\theta_A) \; . \qquad [6]$$

Since $E(\hat{\theta}|\theta) = \theta$ asymptotically (i.e., the maximum likelihood estimator is unbiased for tests of sufficient length) and the covariance between θ and $(\theta-\hat{\theta})$ is zero asymptotically, it may be concluded that $E(\hat{\theta})=E(\theta)$ and

$$V(\hat{\theta}) \simeq V(\theta) + E[V(\hat{\theta}|\theta)] \qquad [7]$$

for tests of sufficient length (Samejima, 1977, pp. 164-166). Also, for tests of sufficient length, $E[V(\hat{\theta}|\theta)]$ can be estimated in large samples with the quantity $M[1/I(\hat{\theta})]$, which is the sample mean of the reciprocal values of the test information function (Birnbaum, 1968, p. 454) evaluated at each examinee's $\hat{\theta}$ (Sympson, 1980).

Thus, an estimate of k_1, corrected for unreliability of the odd and even halves of the linking test, was calculated using

$$\text{est } (k_1) = \left[\frac{\text{est } V(\theta_S)}{\text{est } V(\theta_A)}\right]^{\frac{1}{2}} = \left[\frac{V(\hat{\theta}_S) - M[1/I(\hat{\theta}_S)]}{V(\hat{\theta}_A) - M[1/I(\hat{\theta}_A)]}\right]^{\frac{1}{2}} \qquad [8]$$

$$= \left[\frac{(.710)^2 - .040}{(1.027)^2 - .109}\right]^{\frac{1}{2}} = .701 \; ,$$

where .040 and .109 were the estimates of $E[V(\hat{\theta}_S|\theta_S)]$ and $E[V(\hat{\theta}_A|\theta_A)]$ computed from 511 reciprocal values of each half's test information function. Similarly, k_2 was estimated using

$$\text{est } (k_2) = \text{est } E(\theta_S) - [\text{est}(k_1) \text{ est } E(\theta_A)] \qquad [9]$$

$$= M(\hat{\theta}_S) - [\text{est}(k_1) \; M(\hat{\theta}_A)]$$

$$= (.478) - [(.701)(-.022)] = .494 \; .$$

Since item difficulty (b) and ability (θ) are on the same metric, the function for transforming item difficulty on the SCAT metric (b_S) to the Air Force metric (b_A) is

$$b_A = \frac{b_S - \text{est}(k_2)}{\text{est}(k_1)} \; . \qquad [10]$$

Item discrimination is transformed with $a_A = \text{est}(k_1)a_S$ (Lord & Novick, 1968, p. 381). These transformations were applied to the a and b parameters of all 113 SCAT items in order to have parameters that were expressed on the same metric as the AFHRL items. After parameter transformation, one SCAT item was found to have a discrimination value below .25 and was dropped from the WK pool. Thus, the final WK item pool for the adaptive tests used in this research was comprised of 168 AFHRL items and 112 SCAT items, a total of 280 items.

Table 2 also provides information about the distributions of a, b, and c parameters in the final WK item pool. Items with a < .25 were excluded from the pool. Thirteen items had a equal to the upper bound of 2.0. Ninety percent of the b parameters were between -3.663 and 2.984. This indicates that the WK item pool was suitable for testing individuals with a wider range of abilities than the AR pool (in which 90% of the b parameters fell between -1.533 and 1.876). The item response function (IRF) c parameters for 103 out of 280 items were set to the default value of .155 calculated by LOGIST. Another 43 c parameters were estimated to be .150. The remaining c parameters ranged from .024 to .316. The Pearson product-moment correlation between the a and b parameters in the WK item pool was .488, a value quite close to the correlation observed in the AR item pool.

ASVAB Calibrations

The conventional (non-adaptive) tests administered in this research were Parts 4 and 5 (WK and AR) of Form 7 of the Armed Services Vocational Aptitude Battery (ASVAB). Part 4 of ASVAB-7 contained 30 four-alternative multiple-choice WK items of the same type used in the adaptive-test WK item pool. Part 5 of ASVAB-7 contained 20 four-alternative multiple-choice AR items of the same type used in the adaptive-test AR item pool. These two ASVAB subtests were administered on CRTs along with the experimental adaptive tests. In order to investigate the possible effects of test scoring method on criterion-related validity, each ASVAB subtest was scored three ways. In addition to the traditional number-correct score that is commonly used with the ASVAB, both Bayesian and maximum likelihood IRT ability estimates were generated for each examinee on each ASVAB test.

In order to compute the IRT ability estimates, IRT item parameters were needed. Data were obtained for 2000 Air Force enlistees tested with ASVAB-7 during their basic military training. These data were used in two LOGIST calibration runs in order to generate parameter estimates for the AR and WK ASVAB subtests.

Arithmetic Reasoning. Of 2000 available AR subtest records, 146 were eliminated because the individual either attempted fewer than 10 questions or answered all attempted questions correctly. All 1854 remaining cases attempted the first 10 AR subtest questions, but because some examinees either omitted or did not reach many of the later items, fewer and fewer responses were available for these items. By Item 20, 1572 responses were available. This indicates that the ASVAB-7 AR subtest was somewhat speeded and, as a result, not entirely suitable for calibration with IRT methods (Lord & Novick, 1968, p. 384). Nevertheless, in order to address the question of effects of scoring method on criterion-related validity, the LOGIST calibration of this subtest was undertaken. The resulting IRT parameters are shown in columns 2 through 4 of Table 3.

Table 3
Item Parameter Estimates for Items in ASVAB-7 AR and WK Subtests

Item Number	AR			WK		
	a	b	c	a	b	c
1	.165	-10.481	.250	.741	-4.338	.245
2	.281	-4.037	.250	.494	-4.500	.245
3	.381	-3.527	.250	.735	-2.972	.245
4	.510	-2.727	.250	.186	-7.556	.245
5	.768	-.983	.250	.630	-2.458	.245
6	.642	-1.876	.250	.844	-1.881	.245
7	1.120	-.507	.250	.606	-2.913	.245
8	.515	-1.464	.250	.530	-2.096	.245
9	1.142	-.047	.248	.679	-2.767	.245
10	.955	-.089	.250	.567	-1.881	.245
11	.440	-.961	.250	.697	-1.465	.245
12	.569	-.590	.250	1.241	-1.235	.245
13	.675	.380	.250	.520	-2.170	.245
14	2.000	.197	.250	.888	-1.327	.245
15	.904	.663	.250	1.054	-.870	.245
16	2.000	.352	.183	1.280	-.677	.245
17	.644	.917	.108	1.016	-.671	.245
18	2.000	.522	.293	.753	-1.364	.245
19	.300	-3.325	.250	.639	-.678	.245
20	.905	-.845	.250	.836	-.540	.245
21				.582	-.490	.245
22				.917	.049	.245
23				2.000	-.127	.245
24				2.000	.541	.245
25				1.664	.080	.168
26				.544	-.211	.245
27				.332	.573	.245
28				.959	.748	.245
29				2.000	.907	.240
30				2.000	1.073	.241
Median	.660	-.718	.250	.747	-1.053	.245

The AR item parameters shown in Table 3 indicate that this ASVAB subtest was rather easy for the Air Force enlistee population. In fact, the first three items and Item 19 were so easy that it is doubtful whether LOGIST could generate accurate parameter estimates for these items with data obtained from this population. For items like these, LOGIST is faced with the task of extrapolating virtually the entire IRF from a short, nearly flat, segment of the upper tail of the IRF. Very small changes in the data available in the upper tail can lead to dramatic changes in the numerical values of the a, b, and c parameters obtained for such items. However, such changes will have very little effect on the fit of the estimated IRF to the available calibration data.

Moreover, the obtained parameters can safely be used with future examinees if their true ability levels fall within the range of the original calibration sample.

Word Knowledge. In the calibration of ASVAB WK items, 135 cases were eliminated because all attempted items were correct. The remaining 1865 cases attempted items 1 through 24. After item 24, the number of available responses declined slowly to a low of 1836 responses at Item 30. This indicates that the ASVAB-7 WK subtest was only very slighrly speeded. LOGIST parameter estimates for this subtest are given in columns 5 through 7 of Table 3.

The WK parameters in Table 3 indicate that this subtest was also rather easy for the Air Force enlistee population. Four or five of the items were so easy that their parameters were probably not well estimated. In both this subtest and the AR subtest, most of the items were easy enough to prevent LOGIST from estimating c parameters for these items. The default c value computed by LOGIST in each calibration run was assigned to such items. Since the data used for calibrating the ASVAB subtests were obtained from a large group of individuals from the same population that was sampled for calibration of the AFHRL items in the adaptive test item pools, it may be assumed that the ASVAB item parameters reported here are on approximately the same AR and WK metrics as are the adaptive test items.

Testing Strategies and Scoring Methods

Two different adaptive testing strategies and three methods of scoring the ASVAB subtests were examined in this research. While the adaptive tests were each scored in only one way, in the following discussion the adaptive tests and the various ASVAB scoring approaches will each be referred to as a testing-strategy/scoring-method (TSSM). This usage is consistent with the fact that the strategy adopted for item selection and the method used for scoring an adaptive test are at least partly independent (Sympson, 1975).

Adaptive Tests

Bayesian. One of the adaptive testing strategies investigated was the Bayesian strategy proposed by Owen (1969, 1975) and studied by Jensema (1977), McBride (1977), and Sympson (1977b), among others. In this procedure, it is assumed that the distribution of θ in the population to be tested is a normal distribution with mean and variance that can be specified a priori. In Owen's original development, it was also assumed that the IRF for the correct response on each dichotomously scored item in the item pool conformed to a three-parameter normal-ogive model. The close similarity of the normal-ogive and logistic response models, when the latter includes the scaling constant 1.7 (Birnbaum, 1968, p.399), allows the use of Owen's procedure with items calibrated under the three-parameter logistic model.

Owen derived equations for computing the mean and variance of the posterior distribution of θ, given knowledge of the mean and variance of the normal prior distribution and the response (correct or incorrect) to a single item with known parameters. While the posterior distribution obtained is not itself a normal distribution, Owen proposed approximating the actual posterior distribution with a normal distribution having the same mean and variance. Given this approximation, the same equations can be used again to obtain the mean and variance of the posterior distribution of θ, given the response to a second item with known parameters.

By continuing in this manner, a series of items can be chosen that are adapted to the provisional estimates of θ obtained during the test (the means of the series of posterior distributions) and which are approximately optimal for minimizing the posterior variance of θ at the end of the test. If μ_m ($\underline{m} = 0, 1, 2, \ldots$) is defined as the mean of any prior distribution in the series, and σ_m^2 as the variance of that same prior distribution, then μ_{m+1} and σ_{m+1}^2, the mean and variance of the resulting posterior distribution after the response to item $\underline{m}$+1 with parameters a_g, b_g, and c_g, are given by the following equations:
If item $\underline{m}$+1 is answered correctly,

$$\mu_{m+1} = E(\theta|1) = \mu_m + (1 - c_g) \left[\frac{\sigma_m^2}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}} \right] \left[\frac{\phi(D)}{A} \right] \qquad [11]$$

and

$$\sigma_{m+1}^2 = V(\theta|1) = \sigma_m^2 \left\{ 1 - \left[\frac{1 - c_g}{1 + \frac{1}{a_g^2\,\sigma_m^2}} \right] \left[\frac{\phi(D)}{A} \right] \left[\frac{(1 - c_g)\phi(D)}{A} - D \right] \right\}. \qquad [12]$$

If item $\underline{m}$+1 is answered incorrectly,

$$\mu_{m+1} = E(\theta|0) = \mu_m - \left[\frac{\sigma_m^2}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}} \right] \left[\frac{\phi(D)}{\Phi(D)} \right] \qquad [13]$$

and

$$\sigma_{m+1}^2 = V(\theta|0) = \sigma_m^2 \left\{ 1 - \left[\frac{\phi(D)}{1 + \frac{1}{a_g^2\,\sigma_m^2}} \right] \left[\frac{\frac{\phi(D)}{\Phi(D)} + D}{\Phi(D)} \right] \right\}. \qquad [14]$$

In these equations, $\phi(D)$ is the standard normal density function, $\Phi(D)$ is the standard normal distribution function,

$$D = \frac{b_g - u_m}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}} , \qquad [15]$$

and $A = c_g + [(1 - c_g)\ \Phi(-D)]$ is the marginal probability of a correct response, given the item parameters and a normal prior distribution for θ.

In order to select item m+1 in Owen's procedure, the computer searches the item pool to identify the as-yet-unadministered item that minimizes

$$\underset{u_g}{E}\ [V(\theta|u_g)] = A\ [V(\theta|u_g = 1)] + (1 - A)[V(\theta|u_g = 0)] , \qquad [16]$$

the expectation of the posterior variance of θ given the response u_g to item g. Once the desired item is identified, it is administered and μ_{m+1} and σ^2_{m+1} are computed.

The Owen Bayesian (BAYES) strategy was used to administer a 25-item AR test and a 35-item WK test. However, for many of the subsequent data analyses, the mean of the Bayesian posterior θ distribution after either 20 AR items or 30 WK items was used as the BAYES ability estimate. These test lengths were selected in order to insure comparability of results to the 20-item and 30-item ASVAB AR and WK subtests.

Stratified maximum information. The other adaptive testing strategy used was a stratified maximum information (STMI) strategy. Samejima (1969, p.75) proposed a strategy in which a provisional maximum likelihood estimate of θ would be calculated after each item was administered. At each stage in the test, the next item selected for administration would be the previously unadministered item in the pool which had the largest value of the item information function (Birnbaum, 1968, p. 454) at the current estimated ability level. Item information is defined by

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)Q(\theta)} , \qquad [17]$$

where $P(\theta)$ is the assumed IRF (e.g., Equation 1),
    $Q(\theta)$ is $1 - P(\theta)$,
  and $P'(\theta)$ is the first derivative of $P(\theta)$ evaluated at θ.

During adaptive testing, $\hat{\theta}$ would be substituted for θ in Equation 17. Similar to Owen's procedure, Samejima's proposed strategy would require a complete search of the item pool at each stage of the test in order to identify the optimal item to administer.

To avoid time-consuming pool searches, the STMI strategy "prestratifies" the item pool in terms of item information values at selected levels of θ. For this research, the item information functions of all the items in each adaptive test item pool (AR or WK) were evaluated at θ values ranging from -3.00 to 3.00 in increments of .25. At each selected level of θ, the items in the pool were rank ordered in terms of their information values at that level, and the item numbers of the k = 25 (for the AR pool) or k = 35 (for the WK pool) most informative items were recorded in the order of their information values.

For each item type (AR or WK), the foregoing procedure resulted in 25 ordered vectors of k item numbers, one vector associated with each selected θ level. These vectors were formed into two k-by-25 arrays that were stored in computer memory for rapid access during the administration of the STMI adaptive tests. After each item in a test was administered, a maximum likelihood estimate of ability, $\hat{\theta}$, was calculated using the Fisher scoring algorithm (Kendall & Stuart, 1973, p.51). Then, the item number of the most informative unadministered item in the array column corresponding to the θ value closest to $\hat{\theta}$ was extracted and that item was administered next. Since a given item can be among the k most informative items in an item pool at different levels of θ, a single item often appeared in more than one column of a k-by-25 array of item numbers. As soon as an item was administered, it was removed from all the array columns in which it appeared.

The STMI strategy closely approximates Samejima's proposed strategy over the range of θ in which the item pool has been prestratified. If computer memory provides sufficient storage capacity, the pool can be stratified over a wider range of θ and/or the distance between θ levels at which the pool is stratified can be decreased.

One difficulty with maximum likelihood ability estimation in the context of IRT is that the likelihood function after a single response, or after a series of responses that are all correct or all incorrect, has a maximum at either positive or negative infinity (depending on whether the responses are correct or incorrect, respectively). Thus, unless controls are built into the estimation procedure, the computer will attempt to find this maximum by executing an infinite number of steps along the θ continuum. This same problem can arise with certain item sets and response vectors even when there is a mixture of correct and incorrect responses.

Thus, the following controls were imposed on the numerical procedure (the Fisher scoring algorithm) used to find the maximum of each likelihood function. First, the number of iterations the procedure was allowed to make was restricted to the truncated value of $(i+1)/2$, where i was the number of items administered up to a given point in the test. Consequently, as the test progressed, the number of iterations allowed after items 1, 2, 3, 4, 5, ..., k, was 1, 1, 2, 2, 3, ..., $(k+1)/2$. Second, the largest change in $\hat{\theta}$ allowed in a single iteration was 1.0. Thus, after administering any given item, the largest possible change in

$\hat{\theta}$ was 1.0 times the number of iterations allowed. Third, whenever the change in $\hat{\theta}$ from one iteration to the next was less than .001, the estimate was considered to have "converged." This terminated the numerical iterations and the last value of $\hat{\theta}$ was used as the provisional ability estimate.

Finally, $\hat{\theta}$ was restricted within the interval from -5.00 to 5.00, inclusive. If, during the numerical iterations, $\hat{\theta}$ attempted to go outside this interval, it was set to the appropriate boundary value. If the next iteration stepped $\hat{\theta}$ toward the interior of the bounded interval, the iterations continued until convergence was achieved or the maximum number of iterations allowed at that stage of the test was reached. However, if the next iteration attempted to step $\hat{\theta}$ outside the interval a second time, $\hat{\theta}$ was kept at its assigned boundary value and the iterations were terminated. The boundary value was then used as the provisional ability estimate.

As in the case of BAYES, the STMI strategy was used to administer a 25-item AR test and a 35-item WK test. As before, the values of $\hat{\theta}$ after 20 AR items and 30 WK items were used in most analyses in order to insure comparability with the ASVAB subtests.

ASVAB

Bayesian scoring. Since the ASVAB subtests were not adaptive tests, they were not scored at the time of test administration. Instead, item responses were recorded and the responses were scored three different ways after all the data for this study had been collected. One method of scoring the ASVAB subtests was to generate Bayesian estimates of θ using Owen's equations (Equations 11 to 15). This set of ability estimates will be referred to as the ASVAB/B TSSM.

Scoring of the ASVAB response vectors with Owen's equations was accomplished by processing the items sequentially, as though they were part of an adaptive test. The items were processed in the same order that they appear in ASVAB-7, which was also the order of administration used in this research. The order in which items are processed by Owen's equations is a necessary consideration because the obtained ability estimate will be slightly different if different item orders are used. This characteristic of Owen's procedure is a result of the approximations involved in the method (Sympson, 1977b).

Maximum likelihood scoring. The second method used for scoring the ASVAB subtests was to generate maximum likelihood ability estimates using the numerical procedure described above in connection with the STMI strategy. This will be referred to as the ASVAB/M TSSM. In this case, it was not necessary to process the ASVAB items in a serial fashion as was done with Owen's equations. All 20 AR items or all 30 WK items were treated simultaneously. For each subtest, a single ability estimate, $\hat{\theta}$, based on all the items was generated. The numerical procedure was allowed a maximum of 25 iterations in this case.

Number-correct scoring. The simple number-correct score was also calculated for each ASVAB response vector. This will be referred to as the ASVAB/N TSSM. There were two principal differences between standard ASVAB testing conditions and the conditions established in this research. First, the ASVAB items were administered via CRT rather than in a printed test booklet. This was done to equalize any possible effects of the medium for item administration (Sympson, 1975). This type of effect has been observed in several studies of computer-administered conventional tests (e.g., Sachar & Fletcher, 1978, and references cited therein). Second, the ASVAB subtests, like the adaptive tests, were untimed. As mentioned earlier, there was evidence in the item calibration data that the ASVAB AR subtest is normally given with a time limit that prevents a significant number of examinees from finishing it. In this research, all examinees paced themselves and attempted all items administered. No omitting was allowed.

Adaptive Test Entry Procedures

Fixed Entry

In a conventional (non-adaptive) test, such as the ASVAB subtests used in this research, all examinees take the same items. Thus, the choice of the first item to administer is not a weighty consideration. In adaptive tests, on the other hand, the choice of the first item to administer is usually guided by psychometric considerations. Often, the first item is chosen to be of suitable difficulty for the "average" individual in the examinee population. In view of the scaling established by LOGIST during item calibration, a reasonable approach in the BAYES strategy would be to set $\mu_0 = 0.0$, $\sigma_0^2 = 1.0$, and then to choose the item that minimizes the expectation of the posterior variance. For STMI, a reasonable choice for the first item would be the item which has maximum item information at $\theta = 0.0$. Both of these approaches assume that the individuals to be tested are drawn from the population previously sampled for item calibration and that no additional information related to θ is available about an individual.

In this study it was desired that the adaptive tests function in a manner that would make comparisons with the ASVAB subtests most equitable. Since the ASVAB is designed for use in the type of population encountered at Armed Forces Entrance and Examining Stations (AFEES), it was decided that the adaptive tests should select an initial item that would be appropriate for individuals near the mean of the AFEES population. After this, each adaptive test was allowed to adapt the difficulty of the items to the apparent ability level of each examinee.

Estimating the ability distributions. In order to identify appropriate initial items for the AFEES population, estimates of the mean and standard deviation of AR and WK abilities in this population were needed. Testing of a new sample of individuals from this population was not possible, so a procedure that used existing test data was implemented. Lord and Novick (1968, pp. 387-391) discuss the relationship between the distribution of ability, θ, in a population and the distribution of "true" number- (or proportion-) correct scores on a conventional test for the same population. They show that the non-linear function which transforms θ into true score on a particular test is given by the test characteristic curve (TCC), which is the sum (or average) of the IRFs for correct responses to the items in the test.

Since the relationship between θ and true score is one-to-one in both directions, not only can the TCC be used to convert θ values to true scores, it can also be used to convert true scores to θ values. Thus, if the distribution of true scores on any given test in a selected population is known, the underlying θ distribution for the population can be generated. Note that while the true-score distribution and the TCC are functions of the particular test involved, the resulting θ distribution that is generated is independent of the test used.

In order to generate estimated AR and WK θ distributions for the AFEES population, it was first necessary to obtain an estimated population true-score distribution for an AR and a WK conventional test and to estimate the TCC for each test. Since estimated IRF parameters for subtests AR and WK of ASVAB-7 were already available, estimating the TCC for each of these tests was simply a matter of summing the estimated IRFs for correct responses in each test. The estimated IRFs were generated using the parameters displayed in Table 3.

The estimation of true-score distributions is a complex problem, and only one thorough treatment of the topic has appeared in the psychometric literature (Lord, 1969). Since Lord's procedure for estimating true-score distributions was not available for this research, a simple approximate method was used. First, observed-score distributions from three large AFEES samples (N = 1479, 1387, and 1422) that had taken ASVAB-7 AR and WK subtests were located (Fruchter & Ree, 1977, Tables A-4 and A-5). For each subtest, the three number-correct-score distributions were combined to form an observed-score distribution based on 4308 cases. The mean and standard deviation of the resulting AR distribution were 10.84 and 4.17, respectively. The mean and standard deviation of the WK distribution were 17.03 and 6.92.

Under classical test theory, the linear regression estimate of a true score, given an observed score, is

$$X'_T = [(X_0 - M(X_0)]\rho_{XX} + M(X_0) \ , \qquad [18]$$

where X'_T is the estimated true score, X_0 is the observed score, $M(X_0)$ is the mean of the observed scores, and ρ_{XX} is the test reliability coefficient (Lord & Novick, 1968, p. 65). The standard deviation of estimated true scores computed in this manner, $S(X'_T)$, is equal to $\rho_{XX}S(X_0)$, where $S(X_0)$ is the standard deviation of observed scores. $S(X'_T)$ is smaller than the true-score standard deviation, which equals $\sqrt{\rho_{XX}}\,S(X_0)$ (Lord & Novick, 1968, p. 59). However, $\hat{X}_T$ values, defined by

$$\hat{X}_T = [(X_0 - M(X_0)] \sqrt{\rho_{XX}} + M(X_0) \ , \qquad [19]$$

while not least-squares estimates of X_T, correlate perfectly with X_T' and have a standard deviation equal to that of true scores.

For the purposes of this research, each observed-score-interval boundary ($X_O \pm 1/2$) was regressed toward the observed-score mean by inserting the boundary value in place of X_O, and an estimate of ρ_{XX}, in the formula given above for $\hat{X}_T$. The estimates of ρ_{XX} used in this equation for the two ASVAB subtests were KR-20 coefficients (Lord & Novick, 1968, p. 91) previously calculated in a stratified random sample of 460 AFEES applicants (Jensen, Massey, & Valentine, 1976, Table 5). These coefficients were .84 and .91 for AR and WK, respectively. The regressed boundary values were then projected onto the θ metric via the TCC. The examinees falling within each estimated true-score interval were then assigned to the corresponding (projected) interval on the θ metric. A graphical representation of this procedure for the ASVAB AR subtest is shown in Figure 1.

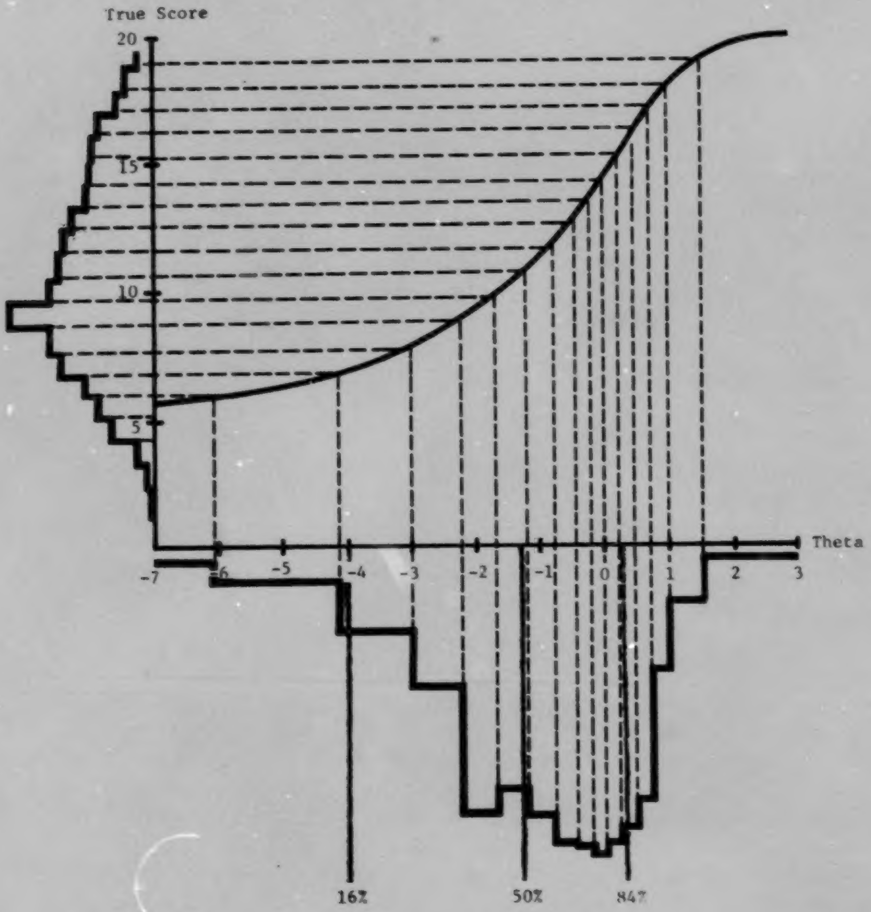
As shown in Figure 1, the lowest θ interval is unbounded at the left and the highest θ interval is unbounded at the right. This prevents computation of the central moments of the estimated θ distribution. In order to approximate the mean and standard deviation of θ, the 16th, 50th, and 84th centiles of the estimated AR and WK θ distributions were obtained by linear interpolation. These centile points are shown in Figure 1 for the AR distribution.

The 16th, 50th, and 84th centiles were selected because the BAYES procedure assumes that the distribution of θ is normal. Unfortunately, as indicated in Figure 1 for AR abilities, the estimated θ distributions of both abilities were clearly negatively skewed, which caused the distance between θ_{16} and θ_{50} to greatly exceed the distance between θ_{50} and θ_{84}. It seems likely that part, though not all, of this negative skew was due to the fact that the IRF lower asymptote parameters (c_g) could not be accurately estimated with the available calibration data.

In general, that portion of a test characteristic curve (TCC) which lies below about $\theta = -2.50$ will not be well estimated with parameters generated by LOGIST, due to the paucity of data at extreme θ levels. The fact that so many estimated true-score intervals fell below the lower asymptote of the estimated TCC of each subtest, and the fact that the average well-estimated c_g value for r-alternative multiple-choice items is usually lower than $1/\underline{r}$ (Lord, 1974, p. 252), suggest that the lower portion of the estimated TCCs obtained in this research were elevated above the true TCCs. If this conjecture is correct, then the true AR and WK θ_{16} values are somewhat higher than the values obtained here. Use of more accurate θ_{16} values would reduce the distance from θ_{16} to θ_{50} somewhat, but the reduced distance would probably still be greater than the distance from θ_{50} to θ_{84}.

Fixed-entry ability estimates. Each examinee was administered a total of six tests. The first three tests were either AR or WK tests,

Figure 1
Transformation of an Approximate ASVAB-7 Arithmetic Reasoning True-Score Distribution into a Theta Distribution via the Test Characteristic Curve



depending on the "order condition" to which an examinee was assigned. The last three tests followed a 5-minute break.

The first three tests a person received were given in what will be referred to as the fixed-entry (FE) condition. In each of these tests, all examinees were administered the same initial item. The choice of an initial item for a test depended on the item type involved and the strategy. For example, for the BAYES AR adaptive test, μ_0, the mean of the initial prior distribution, was set equal to -1.30. This value was chosen because the estimated 50th centile point of the AFEES AR ability distribution was at $\theta = -1.29$. The standard deviation of the BAYES initial prior distribution, σ_0, was set equal to 2.15 because the mean of the estimated θ_{16} to θ_{50} and θ_{50} to θ_{84} distances for the AFEES AR ability distribution was 2.16.

The initial AR prior standard deviation used in this study probably overestimated the actual population value, since the observed θ_{16} to θ_{50} distance was probably inflated. However, it was felt that the use of a "diffuse" prior distribution would be less hazardous to the successful functioning of the BAYES strategy than would an overly "precise" prior distribution. The rationale for this point of view can be seen in considering the limiting case where σ_0 is set equal to a value approaching zero. In this case, no finite amount of data can cause the ability estimate to move away from μ_0, the initial prior mean (see Equations 11 to 15). Thus, all the ability estimates will be (approximately) equal to μ_0 at the end of such a test. The same tendency to restrict the ability estimates to an excessively narrow range around μ_0 will operate to varying degrees whenever σ_0 is underestimated. This would, in most cases, reduce the criterion-related validity of the BAYES strategy.

The initial item for the fixed-entry BAYES AR test was the item in the AR pool that minimized the expectation of the posterior variance given [illegible] = -1.30 and σ_0 = 2.15. The initial item for the fixed-entry STMI AR adaptive test was the item for which I(θ) was highest at θ = -1.25, since this level of θ was closest to -1.29 among the 25 levels at which the AR item pool had been prestratified.

For the estimated WK ability distribution, the obtained values of θ_{16}, θ_{50}, and θ_{84} were -4.30, -1.41, and .38, respectively. Thus, the initial item for the fixed-entry BAYES WK test was that item in the WK item pool that minimized the expectation of the posterior variance given that the values of μ_0 and σ_0 were -1.40 and 2.35, respectively. Similarly, the initial item for the fixed-entry STMI WK test was the item for which I(θ) was highest at θ = -1.50, since -1.50 was closer to -1.41 than any of the other ability levels at which the WK item pool had been prestratified.

Variable Entry

A potential advantage of computerized adaptive testing is the possibility of selecting an initial test item that is tailored to the individual. In order to do this, information about the examinee beyond his/her group membership is needed. One possible source of such additional information is the examinee's performance on a previous test whose scores are correlated with the ability to be estimated (e.g., Brown & Weiss, 1977). In this research, an individualized starting point for each of the last two adaptive tests a person completed was derived from the final ability estimate obtained in the first adaptive test of the same type (BAYES or STMI).

Specifically, if an individual answered WK items in the first half of the testing session, an initial linear regression estimate of the person's true AR ability level (θ_{AR}) was obtained from their estimated WK ability level ($\hat{\theta}_{WK}$). The initial AR estimate ($\tilde{\theta}_{AR}$) was then used to select an appropriate first AR item. This was done separately for the

BAYES and STMI adaptive tests. Thus, the adaptive tests given in the second half of an individual's testing session can be described as having been administered under variable-entry (VE) conditions. Of course, each ASVAB subtest was administered under FE conditions, regardless of which portion of the testing session the subtest appeared in.

In estimating θ_{AR} from $\hat{\theta}_{WK}$, the linear regression equation may be written as

$$\tilde{\theta}_{AR} = \left[\rho^* \frac{S(\theta_{AR})}{S(\hat{\theta}_{WK})}\right] \hat{\theta}_{WK} + E(\theta_{AR}) - \left[\rho^* \frac{S(\theta_{AR})}{S(\hat{\theta}_{WK})}\right] E(\hat{\theta}_{WK}) , \qquad [20]$$

where ρ^* is the correlation between θ_{AR} and $\hat{\theta}_{WK}$. For tests of sufficient length, the value of ρ^* is approximately equal to ρ' / ρ_{AR}, where ρ' is the correlation between $\hat{\theta}_{AR}$ and $\hat{\theta}_{WK}$, and ρ_{AR} is the reliability of $\hat{\theta}_{AR}$ (Lord & Novick, 1968, p. 70). For maximum likelihood estimates of θ obtained from tests of sufficient length, the value of $S(\hat{\theta}_{WK})$ is approximately equal to $S(\theta_{WK})/\sqrt{\rho_{WK}}$, where ρ_{WK} is the reliability of $\hat{\theta}_{WK}$ (Sympson, 1980). For Bayesian minimum-quadratic-loss estimates of θ obtained from tests of sufficient length, the value of $S(\hat{\theta}_{WK})$ is approximately equal to $S(\theta_{WK})\sqrt{\rho_{WK}}$ (Sympson, 1980).

$E(\theta_{AR})$, $E(\hat{\theta}_{WK})$, $S(\theta_{AR})$, and $S(\theta_{WK})$ were set to the values used as prior means and standard deviations in the fixed-entry Bayesian tests. An approximation to the unknown value of ρ' was obtained by calculating the weighted mean of three previously reported correlations between number-correct scores on ASVAB-7 AR and WK subtests (Fruchter & Ree, 1977; Tables 4, 9, and 13). These empirical correlations were based on 1479, 1387, and 1442 cases. The obtained estimate of ρ' was .527.

The estimates of ρ_{AR} and ρ_{WK} that were used were the same values (.84 and .91) used earlier in connection with the creation of approximate true-score distributions for the ASVAB AR and WK subtests. Since these reliability estimates were obtained from non-adaptive, number-correct-scored tests, they presumably underestimated the unknown reliabilities of the adaptive AR and WK tests used in this study.

The procedure described above resulted in the following equations for estimating θ_{AR} from $\hat{\theta}_{WK}$:

for BAYES,

$$\tilde{\theta}_{AR} = .55(\hat{\theta}_{WK}) - .53 \; ; \qquad [21]$$

for STMI,

$$\tilde{\theta}_{AR} = .50(\hat{\theta}_{WK}) - .60 \; . \qquad [22]$$

In a similar manner, the following equations were obtained for estimating θ_{WK} from $\hat{\theta}_{AR}$:

for BAYES,

$$\tilde{\theta}_{WK} = .66(\hat{\theta}_{AR}) - .54 \; ; \qquad [23]$$

for STMI,

$$\tilde{\theta}_{WK} = .55(\hat{\theta}_{AR}) - .68 \; . \qquad [24]$$

The appropriate pair of equations was used to calculate initial ability estimates for starting each examinee's second BAYES and STMI tests. Initial prior standard deviations for second BAYES tests were set equal to the estimated standard error of estimate for predicting one ability from estimates of the other. Specifically, for predicting θ_{AR} from $\hat{\theta}_{WK}$,

$$S(\theta_{AR} - \tilde{\theta}_{AR}) = S(\theta_{AR})\sqrt{1 - (\rho^*)^2} = 1.76 \; . \qquad [25]$$

For predicting θ_{WK} from $\hat{\theta}_{AR}$,

$$S(\theta_{WK} - \tilde{\theta}_{WK}) = S(\theta_{WK})\sqrt{1 - (\rho^*)^2} = 1.96 \; . \qquad [26]$$

In Equations 20 to 26, the value used for ρ^* was $.527/\sqrt{.84}$ when estimating θ_{AR} from $\hat{\theta}_{WK}$ and $.527/\sqrt{.91}$ when estimating θ_{WK} from $\hat{\theta}_{AR}$.

Apparatus

Testing System Hardware

The computerized test delivery system created for this research utilized a Hewlett-Packard 2100S microprogrammable computer with 32K 16-bit words of memory. This central processing unit (CPU) was connected to a Hewlett-Packard 7900A moving-head disc drive, a Texas Instruments 733 KSR printing terminal, and four ADDS Regent 100 CRTs.

The 733 KSR terminal functioned as the test proctor's control console. This unit provided a hard-copy record of all communications between the proctor and the CPU. An important consideration in selecting the control console was the requirement that the printer be virtually silent. This was necessary because the console would be located in the same room as the examinees taking the experimental tests.

The ADDS CRTs were driven at 1200 BAUD (approximately 120 characters per second) and functioned as the delivery and response medium for the six tests administered to each examinee. Communications transmitted to the examinee from the CPU were displayed on a 12-inch-diagonal cathode-ray tube in white-on-black character mode using an 8-by-8 dot matrix for each character. A maximum of 80 characters per line, on up to 24 lines, could be displayed at one time. Examinee responses were entered on a standard teletypewriter (TTY) type keyboard. Examinees were able to review each line of their input for errors and could "erase" the line and re-enter it before "sending" it to the CPU.

Some of the AR items required slight modifications for CRT display. For example, the correct response for one item in the AR item pool was printed in the original test booklet (PE7701) as

$$\frac{300 - 90}{7\tfrac{1}{2} - 2\tfrac{1}{2}} .$$

This option was displayed on the CRT as

$$\frac{300 - 90}{(7\ 1/2) - (2\ 1/2)} ,$$

since the small ½ symbol was not available on the CRTs. It was assumed that these modifications did not alter the difficulty of the items.

Test Administration Software

Exclusive of the Real Time Executive (RTE) operating system provided by Hewlett-Packard, and certain special-purpose assembly-language routines either obtained as part of a Hewlett-Packard program library or written specifically for this research, the testing system implemented for this study was written entirely in FORTRAN IV.

The testing system was made up of a series of self-contained programs and callable subroutines. Each self-contained program could schedule one or more other programs for execution via appropriate instructions addressed to the RTE operating system. The programs served to initialize each testing session, train each examinee in the use of the CRT keyboard, collect biographical information (name and social security number), initialize examinee data files, assign each examinee to one of 12 "order conditions," administer sample items and the six experimental tests in the appropriate order, recognize and deal with invalid examinee responses, keep track of time elapsed in a particular test and examinee response times for each item taken, record the data from the six experimental tests on the examinee data file, terminate testing, debrief each examinee, and close the examinee data file.

System Checkout and Installation

After activation of the integrated hardware/software system, several-hundred trial testing sessions were completed to verify the integrity of the system. The majority of these testing sessions were completed by staff members connected with this research, but in order to test the system and the examinee instructions with naive respondents, approximately 35 volunteers, students in an introductory psychology course, were also tested.

During these "shake-down" sessions, a few instances in which the testing system stopped functioning in the middle of a testing session were noted. Extensive troubleshooting of the system hardware and software failed to reveal the reason for these failures. Since such occurrences were infrequent and the scheduled starting date for testing was imminent, the system was placed in the field in its then current configuration. (During the subsequent period of experimental testing, system failures of this type were experienced six times during 40 days of testing.)

The system was transported to Chanute AFB and two individuals selected by the Air Force were trained as test proctors by project staff. The test proctors activated the testing system each morning, initiated each examinee's testing session, kept a log of the day's testing activities, shut down the system at the end of each day, and periodically changed the removable disc cartridge on which experimental data were accumulating. After only one day of monitored testing activity, these proctor functions were performed without on-site supervision from the research staff.

Subjects

The Experimental Sample

The examinees in this research were 495 individuals enrolled in a technical training course for Air Force Jet Engine Mechanics [Air Force Skill Code (AFSC) 42632]. The examinee sample contained approximately 70% males and 30% females. Most major U.S. racial and ethnic groups were represented, though not necessarily in proportion to their numbers in the general population. With few exceptions, the examinee sample was concentrated in the age range from 18 to 23 years.

The Jet Engine Mechanic (JEM) training course lasted 11 weeks. Each trainee was tested either within a few days before starting the course or sometime during the course. Testing was completed between June 15 and August 11, 1978. While participation in this study was voluntary, trainee participation was strongly encouraged by senior enlisted personnel at the JEM school. Virtually all JEM trainees that were enrolled at the time testing was started and all new trainees arriving while testing was underway volunteered to participate in the study.

Cases Retained for Analyses

While 495 complete test records were obtained in this research, not all were used in the various data analyses. Table 4 summarizes the number of cases remaining after the application of several criteria for case retention. In this table, the total examinee sample is broken down into two groups: individuals who answered WK items during their first three tests and AR items during their final three tests (the WK-AR group), and individuals who answered AR items before WK items (the AR-WK group).

Table 4
Experimental Subgroups Formed by Applying
Alternative Retention Criteria

Sub-group	Retention Criterion	Group WK-AR	Group AR-WK	Group Total
-	None	249	246*	495
1	AR Criterion (A)	245	243	488
2	WK Criterion (W)	247	245	492
3	A + W	244	242	486
4	A + W + Graduation (G)	231	221	452
5	A + W + G + Pre-Enlistment Data	206	200	406

*248 individuals were tested; data from two individuals were lost due to proctor error.

The AR retention criterion and the WK retention criterion were applied to these groups in an attempt to eliminate examinees who were not really trying to do their best during the experimental tests. In the case of AR items, the average examinee response time per item was on the order of 50 to 60 seconds. Individuals with very low average AR response times and low final ability estimates were suspected of not really trying. Consequently, for any analyses involving AR data, individuals whose average AR response time was less than 10 seconds per item were not included. This AR retention criterion affected a total of seven individuals in the original examinee sample, resulting in 488 examinees in Subgroup 1. In the full sample, the regression of mean AR response time on final AR ability estimate was observed to have a positive slope. That is, individuals with high final ability estimates tended to have long average AR response times. None of the seven individuals affected by the AR retention criterion had a high final ability estimate.

In the case of WK items, the situation was not as clear. The regression of mean WK item response time on final WK ability estimate had a negative slope. Individuals with high estimated ability levels tended to have short average WK response times. Thus, a WK retention criterion based on short mean response time alone was not reasonable. Instead, visual examination of four scatterplots of ability estimates by response times was used to identify three individuals whose mean WK response times were distinctly lower than would be predicted from their final WK ability estimates. These "outliers" were not included in subsequent analyses involving WK data, resulting in 492 examinees in Subgroup 2.

In analyses involving both WK and AR data (e.g., correlations between WK and AR ability estimates), the AR and WK retention criteria were applied jointly (Subgroup 3 in Table 4). Since one examinee was affected by both the WK criterion and the AR criterion when they were applied separately, nine individuals were eliminated by the joint application of these criteria.

In analyses involving the experimental criterion measure (end-of-course grade in the JEM course), it was necessary to eliminate examinees who did not successfully complete the JEM course. This graduation criterion, in conjunction with the AR and WK criteria, reduced the total number of cases available for these analyses to 452 (Subgroup 4). The analyses that used this subsample were central to the main purposes of the research study.

Finally, among the 452 cases in Subgroup 4, a total of 406 individuals (Subgroup 5) were identified for whom scores had been recorded on five Air Force pre-enlistment ASVAB composites (Armed Forces Qualifying Test--AFQT, Mechanical, Administrative, General-Technical, and Electronics). In this group, the criterion-related validity of the pre-enlistment ASVAB composites could be compared to the validity of AR+WK composites formed from the experimental test scores.

Data Collection Procedures

The Testing Environment

The computerized testing system was set up in a single, well-lighted, windowless room. CRTs facing in the same direction were separated so that examinees could not readily observe each other's responses.

The testing room was air-conditioned and located immediately adjacent to a classroom and to a large engine maintenance bay in the JEM training school. These three areas were occasionally in use simultaneously and were not separated by closable doors; therefore, some distraction of the experimental examinees by activities in the classroom and the maintenance bay may have occurred from time to time. Control of this type of problem was left to the discretion of the test proctor.

Instructions

A series of instructions appeared on each examinee's CRT screen at various times during the testing session. These instructions were designed to establish and maintain examinee motivation to do well, to show examinees how to enter their names and social security numbers on the CRT keyboard, and to teach examinees how to enter their responses to the test items.

Except for names and social security numbers, and the words "GO" and "STOP" entered at two points in the testing session, the only keyboard responses required of the examinees were the numbers "1" through "5." During an experimental test, the number entered by an examinee corresponded to the particular response alternative the examinee wished to select for the displayed multiple-choice item. The testing system was programmed to accept only the numbers 1 through 4 as responses to four-choice items (all AR items and the ASVAB WK items) and only the numbers 1 through 5 as responses to five-choice items (the adaptive-test WK items).

If, at any time during the instruction or sample-item phases of the testing session, the examinee failed to enter either the specific response requested or the type of response expected by the testing system, the system branched to an appropriate "error screen" and asked the examinee to re-enter the response. If the examinee failed to enter the appropriate response a second time, a message was displayed instructing the examinee to request proctor assistance, and a CRT "beep" tone was sounded to summon the proctor to that testing station. Only the proctors knew the keyboard entry that would cause the system to leave the "proctor call" screen and re-display the screen with which the examinee was having difficulty. During the experimental tests, a similar sequence of error screens was displayed if the examinee entered anything other than a valid numeric response to a test item.

The examinee answered two relatively easy example items prior to starting both the AR and WK portions of the testing session. The second pair of example items was administered, along with appropriate instructions, after the examinee had completed the first three tests and had taken a 5-minute rest break. The final three tests followed immediately.

Test Administration

From the examinee's point of view, it was not obvious that six tests were being administered. After giving instructions and sample items for the item type to be presented in the first half of the testing session (e.g., AR items), all three tests (BAYES, STMI, and ASVAB) in that content domain were presented without interruption. Thus, from the examinee's point of view, it appeared that one long test (e.g., AR) was administered. A similar procedure was used following the 5-minute rest period. An opportunity for the examinee to identify the points of demarcation separating the ASVAB WK test from one or both adaptive WK tests did exist, however, because the ASVAB WK items had four response alternatives and the adaptive test WK items had five response alternatives.

Since examinees were not told that several tests were to be given in each content domain, it seems doubtful that many individuals grasped the significance of the appearance of a block of four-response items. In any case, these items appeared in all three possible (blockwise) sequential locations in the series of WK items.

The experimental test did not impose time limits on the examinees. However, the test instructions did suggest that too much time should not be spent on any one question; in the case of AR items, the instructions indicated that most people take about a minute on each question. Examinees were provided with pencils and scratch paper to aid in completing the calculations associated with the AR items. (The mean time required to complete the entire testing session was calculated for the first 27 experimental examinees. This group averaged 89.1 minutes per session, including the rest break.)

Counterbalancing of orders. Since the use of CRTs as a test delivery and examinee response medium was likely to be a novel experience for most of the individuals participating in this study, and since each testing session was to be rather lengthy, it was anticipated that warmup and fatigue effects might influence examinee performance. In order to distribute such effects evenly over the item types and testing strategies that were studied, each examinee was assigned to one of several order conditions.

Using two different item types (AR and WK) and three testing strategies (BAYES, STMI, and ASVAB), 12 different order conditions were generated. The 12 order conditions are shown in Table 5. One order condition was BAYES WK, STMI WK, ASVAB WK, (break), BAYES AR, STMI AR, ASVAB AR (see WK-AR item-type order in column 1 of Table 5). The other 11 order conditions were generated by reversing the order in which AR and WK items appeared and/or by permuting the order of the three testing strategies. In each order condition, the order of the testing strategies was the same before and after the break.

Table 5
Order Conditions for Test Administration

Item-Type Order and Temporal Position	Strategy Order					
	1	2	3	4	5	6
WK-AR						
1st	BAYES	STMI	BAYES	STMI	ASVAB	ASVAB
2nd	STMI	BAYES	ASVAB	ASVAB	BAYES	STMI
3rd	ASVAB	ASVAB	STMI	BAYES	STMI	BAYES
4th	BAYES	STMI	BAYES	STMI	ASVAB	ASVAB
5th	STMI	BAYES	ASVAB	ASVAB	BAYES	STMI
6th	ASVAB	ASVAB	STMI	BAYES	STMI	BAYES
AR-WK						
1st	BAYES	STMI	BAYES	STMI	ASVAB	ASVAB
2nd	STMI	BAYES	ASVAB	ASVAB	BAYES	STMI
3rd	ASVAB	ASVAB	STMI	BAYES	STMI	BAYES
4th	BAYES	STMI	BAYES	STMI	ASVAB	ASVAB
5th	STMI	BAYES	ASVAB	ASVAB	BAYES	STMI
6th	ASVAB	ASVAB	STMI	BAYES	STMI	BAYES

The 12 order conditions were assigned to examinees in a systematically rotated fashion in order to insure that approximately the same number of individuals was tested under each order condition. Exclusive of two cases that were lost due to proctor errors, 40 cases were tested under one order condition, 41 cases were tested under each of eight other order conditions, 42 cases were tested under each of two other order conditions, and 43 cases were tested under the remaining order condition. For many of the analyses, individuals tested under the six WK-AR order conditions were combined and individuals tested under the six AR-WK order conditions were combined, thus creating two large groups with sample sizes of more than 200 in each group.

Within the large WK-AR group and the large AR-WK group, warmup and fatigue effects were not evenly distributed over the two item types. However, such effects were evenly distributed over testing strategies since each strategy was administered to approximately one-third of the individuals in each group at each of the six possible temporal positions in the series of tests. In those analyses that combined the large WK-AR and AR-WK groups (e.g., certain criterion-related validity analyses), warmup and fatigue effects should be evenly distributed over both item types and testing strategies. As mentioned earlier, adaptive tests given before the break were administered under a fixed-entry (FE) condition and adaptive tests given after the break were administered under a variable-entry (VE) condition. Thus, in this study, warm-up and fatigue effects, if present, are necessarily confounded with the effects of adaptive-test entry type.

Item "fill-in" procedure. In order to make all the items in the WK and AR adaptive-test item pools available to both adaptive testing strategies (BAYES and STMI), an item "fill-in" procedure was implemented. In this procedure, whichever adaptive testing strategy came first in an examinee's assigned order condition was allowed to select and administer any items it required from the appropriate item pool. The adaptive testing strategy that followed then checked to see whether an item it required had been administered during the first adaptive test. Whenever an item had been administered in the first adaptive test, the previously recorded item response was used without displaying the item on the CRT screen a second time. Whenever an item had not been previously administered, it was displayed on the CRT screen and the examinee response was recorded. This procedure continued throughout the second adaptive test. Thus, those items selected by the second adaptive test, but which had not been previously administered, were "filled in" to complete the item response vector for that test.

The same procedure was used for the two adaptive tests that followed the 5-minute rest period. The strategy used in an examinee's third adaptive test was allowed to select and administer any items it required from the appropriate item pool. Then the strategy used in the fourth adaptive test "filled in" any items it selected that had not been administered during the third adaptive test.

The item fill-in procedure served three purposes. First, to the extent that the two adaptive testing strategies selected the same items, significant amounts of examinee time could be saved. Because each testing session was to be rather lengthy, this offered the possibility of reducing fatigue effects and controlling the problem of declining examinee motivation. Second, the very question of a tendency for the adaptive strategies to select the same items was of psychometric interest. Even though the BAYES and STMI strategies use different criteria for selecting items, it was known that these criteria are not completely independent of one another. The fill-in procedure allowed direct assessment of the tendency for the BAYES and STMI strategies to select the same items. Finally, the fill-in procedure insured that both adaptive strategies had access to the best items in the two adaptive-test item pools. If the strategies had been forced to use different items, observed differences among the test validities and other dependent variables could have been interpreted as effects that would diminish or disappear if each strategy had been given access to the entire item pool.

Use of an item fill-in procedure of the type implemented in this research will cause errors of measurement for the adaptive testing strategies to be correlated. If the difference between an examinee's ability estimate obtained under one strategy and the mean of ability estimates for that strategy among people at the examinee's same (true) ability level is positive, then the analogous difference obtained in the other adaptive test will also tend to be positive. Conversely, if one difference is negative, the other will tend to be negative. Correlated errors of measurement of this type were not believed to be detrimental to the purposes of the research. While correlations between ability estimates obtained under the BAYES and STMI strategies were higher in this research than they would have been if the strategies had selected items from parallel, but independent, item pools, the correlation between WK and AR ability estimates from a given strategy, the correlations between ability estimates from a given adaptive strategy and scores derived from the ASVAB, and the criterion-related validity coefficients for a given strategy, were all exactly what they would have been if only that strategy had access to the item pools. This would not be the case if the two strategies had selected items from two independent (and only approximately parallel) item pools or had alternated in selecting items from a single item pool.

The Criterion

The criterion measure used in the validity analysis portion of this study was end-of-course grade in the JEM training school. Criterion scores were available for 461 individuals who had been tested with the experimental tests and who subsequently completed the JEM course.

The training course was broken down into four parts, or "blocks." At the end of each block of instruction, trainees were administered a performance examination and a written examination covering the material taught in the block. The performance examinations required each trainee to demonstrate proficiency with respect to specific job-related tasks spelled out in the course objectives. Trainee proficiency was rated as satisfactory or unsatisfactory. A rating of satisfactory on an end-of-block performance examination was required before a trainee could attempt the written examination for the block.

The end-of-block written examinations were comprised of multiple-choice items, each scored correct or incorrect. The written examinations for Blocks 1, 2, 3, and 4 contained 50, 50, 50, and 30 items, respectively, at the time data for this research were collected. A trainee received 2 points for each item answered correctly in the first three written examinations and 3 1/3 points for each correct answer in the Block 4 written examination. Minimum passing scores on the written examinations were 70, 72, 68, and 60 for Blocks 1 through 4 respectively. An individual who failed to achieve a passing score on an end-of-block written examination was allowed to take an alternate form of the examination at a later date. If the person passed the second test, the written-test score for that block was set equal to the minimum written-test passing score for the block, regardless of the number of points earned in the second test.

The end-of-course grade assigned to an individual by the JEM school was equal to the mean of the four written-test scores that the individual had earned. The lowest end-of-course grade observed among the 461 individuals who completed the course successfully (i.e., attained a rating of satisfactory on all four performance examinations and passing scores on the written examinations) was 68. The highest course grade observed was 99. These extreme values span virtually the entire range of possible passing scores. Among the 452 individuals ultimately used in the criterion-related-validity-analysis portion of this research (Subgroup 4), the mean and the standard deviation of the criterion scores were 83.73 and 6.97, respectively.

It cannot be assumed that the available criterion scores were highly reliable. In addition to the inevitable problem of measurement error when trying to estimate a person's "true" achievement level in the JEM course, there was also a problem with inaccurate reporting of examinees' block scores. Anomalous data were noted in the scores reported for several of the examinees in Subgroup 4. Consequently, a request for verification of all scores was issued and another set of data records was received. Analyses proceeded using the verified data. Unfortunately, evidence for the existence of at least one remaining transcription error was later uncovered. It was found that one individual's Block 3 written test score had been reported as 60 when the minimum passing score for that block was 68. Subsequent investigation revealed that the Block 3 score should have been reported as 70. Fortunately, the value of the end-of-course grade that was used in the validity analyses was only two points lower than it should have been as a result of this error.

It is possible that other clerical errors of this type remained undetected. While such errors will tend to reduce the criterion-related validity of all the TSSM combinations studied, they should not vitiate comparisons among the various strategies.

Design

Independent Variables

Analyses for single tests. The primary objective of this research was to investigate the effects of the five TSSM combinations (BAYES, STMI, ASVAB/B, ASVAB/M, and ASVAB/N) on criterion-related validity. In the analyses for single tests, TSSM was treated as a nominal independent variable with five "treatment levels" that were fully crossed with the two treatment levels (WK and AR) of an Item Type independent variable, and the two treatment levels (first half of testing session and second half of testing session) of an Order independent variable. Since test scores were obtained for each examinee under every level of these three independent variables, they may be referred to as "within subjects" independent variables. (However, as noted below, a given examinee was not tested under every possible combination of levels of the independent variables.)

These considerations suggested that a useful way to conceptualize the data structure in the analyses for single tests was in terms of a 5 x 2 x 2 three-way cross-classification with repeated observations on each factor. Within each of the 20 cells of this cross-classification, the test scores of approximately 246 examinees were obtained. Approximately 41 of the examinees in each cell were tested under each of the six possible order permutations of BAYES, STMI, and ASVAB. This served to counterbalance any "micro" order effects that might exist within each half of the testing session.

As noted earlier, adaptive tests administered before the 5-minute break were given under fixed-entry (FE) conditions, and adaptive tests administered after the break were given under variable-entry (VE) conditions. Thus, any effects associated with adaptive-test entry-type were confounded with Order in the analyses for single tests. The confounding of Order and entry-type effects was not complete, however, since the ASVAB subtests were administered under FE conditions regardless of the Order condition.

By treating Item Type and Order as independent variables, the effects of these two variables could be statistically controlled during the analysis of single-test validity coefficients. This provided more precise statistical tests of the effect of TSSM, the independent variable of primary interest. The decision to treat TSSM, Item Type, and Order as within-subjects variables was prompted by a desire to maximize the sensitivity of the analyses conducted to detect the effects of these variables on criterion-related validity.

Since each examinee was "nested" within one of two Item-Type-by-Order conditions (either WK-AR or AR-WK), any test of the two-way interaction between Item Type and Order conducted in the analysis of single-test validities had to be based on a contrast between independent groups of examinees. In most cases, such tests will be less sensitive than statistical tests based on repeated observations of the same examinees. It was necessary to treat Item-Type-by-Order as a "between-subjects" variable in the analyses for single tests because of practical limitations on the length of time during which examinee motivation could be maintained at high levels and because of the unavailability of parallel item pools to use with each examinee under the two Item-Type-by-Order conditions. This situation seemed acceptable, since there was no a priori reason to anticipate the existence of such an interaction.

Analyses for composites. The criterion-related validity of composite test scores (AR scores combined with WK scores) was also studied. In these analyses, Item Type (AR versus WK) was no longer treated as an independent variable. Moreover, Order (first versus second half of testing session) was replaced by Content Order (AR test administered before WK test or vice versa) as an independent variable. Thus, a suitable conceptualization of the data structure for the analyses of composite-test validities is a two-way cross-classification with five levels on a within-subjects factor (TSSM) and two levels on a between-subjects factor (Content Order). As will be discussed below, the Content Order independent variable in the analyses for composite scores was logically related to the Item-Type-by-Order two-way interaction described above in connection with the three-way cross-classification for single-test validities.

Dependent Variables

In the analyses for single tests, four different dependent variables were examined. The dependent variables were test score, mean examinee response time, mean computer response time, and criterion-related validity.

Test score. This dependent variable was the value of either the IRT ability estimate or the number-correct score generated under a particular combination of levels of the three independent variables. While IRT ability estimates derived from different tests under different conditions, but using items calibrated on the same metric, are directly comparable, comparisons of IRT ability estimates with number-correct scores, and comparisons among number-correct scores from different (nonparallel) tests, are not meaningful. These considerations guided the analysis of ability estimates and number-correct scores. In all comparisons between ability estimates obtained from the two adaptive tests and ability estimates obtained from ASVAB, the adaptive-test ability estimates were based on 30 and 20 items for WK and AR, respectively, in order to make the comparisons equitable.

Mean examinee response time. This dependent variable was computed for a given examinee and a given test as the mean elapsed time in seconds, over all items in the test, from the beginning of item presentation until the examinee responded to each item. Thus, each examinee had six mean examinee response time (MERT) scores--one each for BAYES, STMI, and ASVAB under both WK and AR conditions. Whenever the second adaptive test administered during either the WK or AR half of a testing session selected an item that had already been administered by the first adaptive test in that half, not only was the original item response utilized by the second test, but the original examinee response time was recorded for the second test as well. Thus, to the extent that both adaptive testing strategies used the same items, a strong correlation, above and beyond that due to individual differences in response rate, was induced between MERT values for the two adaptive tests. This effect was not a problem, given the objectives of the research, since the expectation of the MERT values for each adaptive test would still be equal to the value that would be obtained if only one adaptive test had been administered (assuming that warmup and fatigue effects have been properly counterbalanced).

Mean computer response time. This dependent variable was computed for a given examinee and a given test as the total time elapsed from the start of the test (exclusive of instructions and sample questions that were presented before the start of the first test that used a given item type) to the end of the same test, minus the sum of the examinee response times for that test, divided by the number of items administered. Unfortunately, this method of determining average computer response time turned out to have one serious drawback. The clock measuring examinee response time was shut off and reset to zero whenever the examinee responded, regardless of whether the response was valid (admissible) or invalid. Due to an oversight during system development, a separate record was not kept of the elapsed time between examinee entry of an invalid response and entry of a valid response. Thus, since the clock measuring elapsed time for the entire test continued running after an invalid response, the value of the mean computer response time (MCRT) variable was inflated for a given test whenever the examinee entered one or more invalid responses during that test. In a few cases where the examinee required proctor assistance but the proctor was busy helping other examinees, a significant amount of waiting time elapsed that was ultimately, and irretrievably, confounded with actual computer response time. These problems were considered in conducting the analysis of MCRT values.

MCRT was computed only for the first of the two adaptive tests that a person received during each half of the testing session. This was because the second adaptive test in each half utilized the item fill-in procedure described earlier. Under the item fill-in procedure, normal computer response time was increased by the need to check whether each item selected had been administered during the previous adaptive test. On the other hand, computer response time was reduced somewhat whenever it was found that an item had previously been administered, since there was then no need to display it.

In view of these considerations, the data from those examinees assigned to order conditions in which BAYES preceded STMI were used to compute MCRT values for BAYES. Similarly, data from examinees assigned to order conditions in which STMI preceded BAYES were used to compute

MCRT values for STMI. Computer response times under AR and WK conditions were kept separate, but within each item type data from both content orders (WK-AR and AR-WK) were pooled for the data analyses.

As it turned out, after eliminating three examinees due to the WK retention criterion, a total of 246 examinees who had taken BAYES WK before STMI WK and 246 who had taken STMI WK before BAYES WK remained. Elimination of seven cases due to the AR retention criterion gave 244 cases in each of these groups. Since ASVAB was not subject to the item fill-in procedure, computer response times for the ASVAB tests were available from members of both groups. Thus, a total of 492 ASVAB WK and 488 ASVAB AR MCRT values were analyzed.

Single-test validity. The last dependent variable studied in the analyses for single tests, and the one of primary importance to the objectives of this research, was the level of criterion-related validity observed under each possible combination of levels of the independent variables. Criterion-related validity was indexed by the Pearson product-moment correlation between the ability estimates or number-correct scores obtained under a particular combination of independent variable levels and the criterion scores (end-of-course grades) of the examinees observed under that set of conditions. Since each examinee was nested within an Item-Type-by-Order condition (either WK-AR or AR-WK), each examinee contributed to 10 of the 20 validity coefficients computed for single WK or AR tests.

Composite-test validity. In addition to the analysis of criterion-related validity coefficients for single tests, the criterion-related validity of composite scores (linear combinations of WK and AR θ estimates and linear combinations of WK and AR number-correct scores) was also studied. Both equally-weighted and optimally-weighted (least-squares) composites of WK and AR test scores were studied. In both cases, the formation of composite scores reduced the total number of validity coefficients from 20 to 10 and resulted in a two-way cross-classification with five levels on the first factor (TSSM) and two levels on the second factor (Content Order; i.e., WK-AR or AR-WK). Data from each examinee were involved in 5 of the 10 validity coefficients computed for equally-weighted composites and 5 of the 10 validity coefficients computed for optimally-weighted composites.

DATA ANALYSIS PROCEDURES

The Adaptive Test Fill-In Procedure

As mentioned earlier, a question of psychometric interest that has not been investigated previously is the tendency for different adaptive testing strategies to select the same items. The item fill-in procedure used in this research allowed a direct assessment of this tendency for two strategies that represent the state of the art in adaptive testing.

To address this issue, the number of items "filled in" (i.e., actually administered) during each examinee's second adaptive WK test and during each examinee's second adaptive AR test was determined among the 452 examinees that made up Subgroup 4. While data for WK and AR adaptive tests were treated separately, results for the two strategies (BAYES and STMI) were pooled. Relative frequency distributions of the number of items filled in for each item type were examined.

Characteristics of Ability Estimates

Out-of-Bounds and Non-Converged Maximum Likelihood Estimates

One aspect of the analysis of ability estimates obtained in this research was an examination of "out-of-bounds" and "non-converged" maximum likelihood estimates of θ. First, the number of "out-of-bounds" ability estimates was determined for the STMI AR and WK tests and the ASVAB/M AR and WK tests. Out-of-bounds estimates had been set to either -5.0 or 5.0, so the number of cases falling at either of these boundaries was determined. These counts were determined separately for the WK-AR and AR-WK groups within Subgroups 1 and 2 at test lengths of 20 and 30 items for STMI AR and WK, respectively, in order to facilitate comparison with the corresponding ASVAB subtests.

Since the STMI strategy had generated an ability estimate after each AR or WK item, the number of out-of-bounds estimates obtained after administering each item was also determined. This count was made in Subgroup 4 for items 1 through 25 in STMI AR and items 1 through 35 in STMI WK.

The number of "non-converged" cases after administering each item in STMI AR and STMI WK to members of Subgroup 4 was also determined. Non-convergence implies that the change in $\hat{\theta}$ in the last allowed numerical iteration was greater than .001. The number of non-converged estimates obtained in scoring ASVAB/M AR and ASVAB/M WK was also determined in this subgroup.

Distributions and Correlations

The relative frequency distribution, the mean, standard deviation, skew, kurtosis, minimum, and maximum of each estimator or score were also determined. Frequency distributions and summary statistics were obtained for the WK-AR and AR-WK groups separately within Subgroups 1 and 2. This was done for all 10 combinations of TSSM and Item Type.

Bivariate scatterplots were generated for each of the 10 possible pairings of either two ability estimates or one ability estimate and an ASVAB number-correct score for the AR and WK item types. This was done within the Subgroup 1 and 2 WK-AR and AR-WK groups separately, giving a total of 40 scatterplots. Examination of these plots did not suggest the presence of any strongly non-linear regressions, although there was a slight tendency toward non-linearity at extreme values of the maximum likelihood estimator. Since the observed relationships were essentially linear, Pearson product-moment correlations were computed.

Bivariate correlations were also computed within and between the WK and AR domains, separately for the Subgroup 4 WK-AR and AR-WK groups. The correlations in this summary table were computed using modified boundaries for the maximum likelihood ability estimates from STMI and ASVAB/M. These modified boundaries, which will be described below, were adopted following an examination of the effects of limiting boundary values on the criterion-related validity of STMI and ASVAB/M ability estimates.

Information

In order to determine how well each TSSM could discriminate individuals at a given level of ability from individuals at nearby ability levels, score information functions (Birnbaum, 1968, p. 453) were generated for each TSSM. Score information functions for the ASVAB/N subtests were computed analytically (Lord, 1980, p. 73). Score information functions for fixed-entry BAYES and STMI, ASVAB/B, and ASVAB/M were estimated from computer simulations in which a large number of simulated examinations were administered at each of a large number of θ levels.

In comparing the adaptive-test information functions to the ASVAB/N information functions, the ASVAB functions were scaled up to represent subtests of the same length as the adaptive tests (Lord, 1970, p. 155). The adaptive-test information functions were also compared to curves obtained by evaluating the amount of information available from the most informative items in each item pool at several θ levels.

Response Times

Mean Examinee Response Time

Six frequency distributions--one each for BAYES, STMI, and ASVAB under both WK and AR conditions--were constructed for the WK-AR and AR-WK groups separately in Subgroups 1 and 2. In addition, the mean, standard deviation, skew, kurtosis, minimum value, and maximum value of each of these MERT distributions were determined.

Mean Computer Response Time

Data for the analysis of MCRT for BAYES were obtained from individuals who had taken BAYES before STMI; MCRT data for STMI were obtained from individuals who had taken STMI before BAYES. Frequency distributions and summary statistics were obtained for each of these groups on the appropriate AR and WK adaptive tests and also on the ASVAB AR and WK tests.

Since many of the larger MCRT values in each distribution were probably contaminated, summary statistics that would be strongly influenced by extreme values were not computed. Instead, the mode, the minimum value, and the 25th, 50th, 75th, and 90th percentiles of each distribution were determined. The mode of each grouped frequency distribution was computed using quadratic interpolation between the three largest observed relative frequencies in the distribution. The selected centile points of the distributions were computed using linear interpolation within corresponding cumulative relative frequency distributions.

Evaluation of Variable Entry Procedure

As described previously, the adaptive tests administered before each examinee's 5-minute break were given under fixed-entry (FE) conditions and the adaptive tests administered after the break were given under variable-entry (VE) conditions. Presumably, if the initial ability estimates used in starting a VE test correlate substantially with underlying true ability, the first several items administered under VE conditions will be more appropriate for an examinee than items administered under FE conditions. This should result in more accurate ability estimates, particularly in the early stages of the test.

If each examinee's true ability level were known, a plot of the correlation between estimated and true ability as a function of number of items administered would show a negatively accelerating curve that asymptotes toward 1.0 as test length increases without limit. Since true abilities were unknown, interim ability estimates were correlated with the final ability estimates obtained at the end of each adaptive test. It was assumed that any advantage inherent in the VE procedure would manifest itself in higher correlations between interim and final ability estimates, particularly in the early stages of a test.

Validity Analyses

Single Tests

Sequential validity analysis. Pearson product-moment correlations were computed between provisional ability estimates and criterion scores at each stage of each adaptive test. These "sequential validities" were plotted for the WK-AR group from Subgroup 4 for BAYES WK and STMI WK, and compared to the criterion-related validity of ASVAB/N WK in this group at 30 items. Similar correlations were plotted for the AR-WK group in Subgroup 4 for BAYES AR and STMI AR and compared to the validity of ASVAB/N AR at 20 items.

The AR-WK group was not used in plotting the sequential validity data for WK adaptive tests because the variable-entry (VE) procedure induced an a priori correlation between the initial WK ability estimates and criterion scores. This a priori correlation was equal to the final value of the criterion-related validity coefficient for the AR adaptive test of the same type and occurred because the initial WK ability estimate generated under VE conditions was a linear transformation of the final AR ability estimate obtained under FE conditions. Similar considerations indicated that the WK-AR group should not be used in plotting the sequential validity data for AR adaptive tests. To do so would confound final WK validity obtained under FE conditions with sequential AR validity under VE conditions.

Multivariate linear-model analysis. The sequential validity analysis helped to clarify the effect of adaptive-test length on criterion-related validity. However, it did not provide a procedure for making statistical inferences about the effects of TSSM, Item Type, and Order on the level of single-test validity. While such inferences might be approached through a series of statistical tests of the significance of differences between selected pairs of validity coefficients (Glass & Stanley, 1970, pp. 311, 313), interpretation of the resulting large number of correlated significance tests would be difficult and the familywise error rate (Kirk, 1968, p. 85) for each major hypothesis of interest would be well above acceptable levels.

Since data had been collected in a manner congruent with a three-way analysis of variance (ANOVA) layout in the case of single-test validities, an "ANOVA-like" analysis of the obtained validity coefficients was indicated. It was clearly desirable to assess the "main effect" of each independent variable on the level of test validity while the effects of the other independent variables were statistically controlled. Also, the possibility of testing for the presence of interactions among the independent variables was attractive.

To conduct a univariate ANOVA (with validity coefficients as response measures) would require the following three conditions to be met: (a) the response measure (the validity coefficient in each cell of the three-way cross-classification) would have to be approximately normally distributed over random samples; (b) the sampling variance of the response measure would have to be approximately constant over treatment combinations; and (c) the variance-covariance matrix among the sample validities would have to satisfy certain rather restrictive structural requirements (Winer, 1971, pp. 281-283, chap. 7).

If sample size is moderate, the first condition above will not be satisfied unless the population validities are all near zero. If sample size is large, the sampling distribution of a correlation coefficient will be approximately normal, but the closeness of the approximation will depend on both the value of the population correlation (ρ) and the sample size (N). The larger the absolute value of ρ, the larger N must be in order to insure approximate normality.

The variance of the sampling distribution of a correlation coefficient also depends on both the value of ρ and the sample size. However, if the hypothesis of equality of population validities under all possible treatment combinations is true, and if all the sample validities have been computed using approximately the same number of observations, the sampling variances of the validities will be approximately equal. On the other hand, if any of the statistical tests associated with an

ANOVA of (raw) validities were significant, it would suggest that the homoscedasticity assumption must be false for some of the cells in the layout and that, in turn, the p-values associated with the non-significant tests were probably inaccurate.

The normality and homoscedasticity requirements listed above could be satisfied by conducting an ANOVA that used Fisher's transformation (Kendall & Stuart, 1973, pp. 304-305) of the sample validities as the response measure. For moderate to large samples from a bivariate normal population, this transformation provides a sample statistic (referred to here as $\underline{z}^*$) that is approximately normally distributed regardless of the value of the population correlation. Moreover, the sampling variance of this statistic is (approximately) 1/(N-3), a quantity that is also independent of the population correlation. However, one problem remains. An analysis of $\underline{z}^*$ values would not insure that the structural requirements imposed on the variance-covariance matrix among the $\underline{z}^*$ would be satisfied. In fact, due to widely varying levels of dependency among the various ability estimates and/or scores computed in this research, it seemed that making any a priori assumptions about the structure of the variance-covariance matrix among the $\underline{z}^*$ would be hazardous at best.

When the structural requirements imposed on a variance-covariance matrix among dependent sample means cannot be satisfied in a traditional ANOVA, one of two approaches is usually followed. Either an approximate, but conservative, test procedure is implemented, with an attendant loss of statistical power, or a multivariate ANOVA (MANOVA) is conducted (see Collier & Hummel, 1977, pp. 158-173, 211-233). In the MANOVA approach, repeated observations of a single response measure are treated as though they were individual observations taken from several correlated response variables. In view of the fact that most hypothesis tests regarding differences between validity coefficients tend to have rather low power (Lord, 1978, pp. 426-427), approximate procedures that would further reduce statistical power were considered ill-advised in the context of this research. Thus, a "MANOVA-like" analysis was indicated.

A detailed discussion of the rationale behind the statistical inference procedure that was developed for the analysis of test validity coefficients in this research is given elsewhere (Sympson, in preparation). The principal assumptions of the procedure and an overview of the methodology will be given here, but rigorous arguments and supporting proofs that provide formal justification for the procedure will be omitted. This inference procedure may be characterized as a multivariate linear-model analysis of test validity coefficients.

As is widely known, fixed-effects ANOVA can be accomplished through the mechanism of univariate multiple linear regression with "coded" independent variables (Cohen & Cohen, 1975, pp. 3-5, chap. 5). Such analyses are special cases of the univariate general linear model. In these analyses, the main effect of each nominal independent variable in the ANOVA layout is associated with a particular set of coded independent variables in the linear (multiple regression) model. Each independent variable in the ANOVA layout contributes one fewer coded variables to the main-effect portion of the linear model than there are "treatment levels" on that variable.

Interactions among the independent variables in the ANOVA layout are also represented by sets of coded independent variables in the linear model. The set of coded variables representing the interaction of any two nominal independent variables in the ANOVA is obtained by taking the element-by-element set product of the two sets of coded variables representing their main effects. Thus, if there are f treatment levels associated with nominal independent variable F, and g treatment levels associated with nominal independent variable G, their two-way interaction, FG, will be represented by a set of $(f-1)(g-1)$ coded variables in the interaction portion of the linear model. The kth element of a coded FG two-way interaction variable will be equal to the product of the kth element of one particular F main-effect variable and the kth element of one particular G main-effect variable.

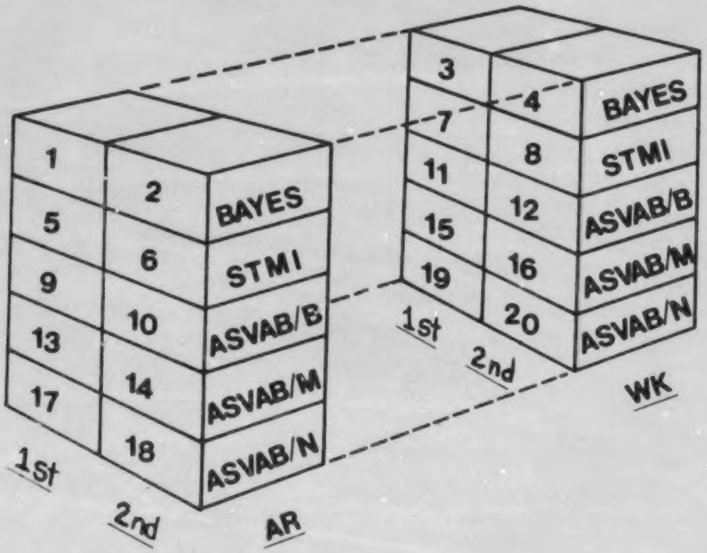
Higher order interactions are represented in the same manner. For example, the three-way interaction of the two previous variables with a nominal independent variable H, having h treatment levels, would be represented by $(f-1)(g-1)(h-1)$ coded variables in the interaction portion of the linear model. These coded variables would be obtained by generating the element-by-element set product of the FG two-way interaction variables described above with the coded main-effect variables for variable H.

Testing main effects or interactions in a linear-model approach to ANOVA is accomplished by testing whether inclusion in the linear model of the entire *set* of coded variables associated with some particular main effect or interaction significantly increases the squared multiple correlation for predicting the response measure (Cohen & Cohen, 1975, pp. 141-144). If the multiple correlation does not increase significantly when a particular set of coded variables is added to the model, the main effect or interaction associated with that set of variables is declared non-significant in the ANOVA.

The significance tests associated with a standard multiple linear regression analysis of ANOVA data require the same three conditions (normality, homoscedasticity, and patterning of the variance-covariance matrix among response measures) that were specified earlier (Cohen & Cohen, 1975, pp. 48-49, p. 404). Since the required variance-covariance structure cannot safely be assumed for the type of data collected in this research, a univariate linear-model analysis of the sample z^* values could not be justified. Instead, a multivariate extension of the linear-model approach was implemented. This procedure is conceptually similar to a MANOVA of correlated sample means by the general linear model (Bock, 1[illegible], chap. 7).

Cell coding. For the analysis of single-test validities, the various treatment combinations associated with the three independent vari-

Figure 2
Cell Codes in the 3-Way Cross-Classification for Single-Test Validities



ables were assigned numerical "cell codes." These codes are shown in Figure 2. Each of the 20 TSSM validity coefficients computed for single AR and WK tests was associated with a particular combination of treatment levels and, thus, a particular cell code. For example, the criterion-related validity coefficient for the BAYES AR adaptive test when administered in the first half of an individual's testing session was associated with cell code 1. Similarly, the validity coefficient for the ASVAB/N WK test among individuals who took this test during the second half of their testing session was associated with cell code 20. The validity coefficients associated with cell codes 1, 4, 5, 8, ..., 17, and 20 were computed on one group of individuals (the AR-WK group in Subgroup 4) while the validity coefficients associated with cell codes 2, 3, 6, 7, ..., 18, and 19 were computed on another group of individuals (the WK-AR group in Subgroup 4).

The first step in the statistical analysis of single TSSM validities was to array the 20 obtained criterion-related validity coefficients in a row vector, $\underset{\sim}{v}$, such that the first element was the validity coefficient associated with cell code 1, the second element was the validity coefficient associated with cell code 2, and so on through the 20th element. A similar ordered vector, $\underset{\sim}{z}$, was created in which the 20 elements were the Fisher $\underline{z}$* transformations of the criterion-related validity coefficients in $\underset{\sim}{v}$. Next, a complete set of 20-element coded vectors for a linear-model analysis of a 3-way cross-classification was generated. There were four main-effect vectors for TSSM, one main-effect vector for Item Type, one main-effect vector for Order, four interaction vectors for TSSM by Item Type, four interaction vectors for TSSM by Order, one interaction vector for Item Type by Order, and four interaction vectors for TSSM by Item Type by Order.

While a variety of coding schemes could have been used in creating the set of 19 vectors for the linear-model analysis, the effects coding method (Cohen & Cohen, 1975, pp. 188-190) was selected. As noted by Cohen & Cohen (p. 189), the choice of a particular coding scheme in a linear-model analysis does not influence the conclusion that is reached regarding the statistical significance of each main effect or interaction. However, for this research, effects coding provided a particularly convenient method for conducting desired one-degree-of-freedom a posteriori significance tests (Kirk, 1968, pp. 87) following the detection of a significant main effect or interaction. The 19 coded vectors used in this research are shown in Table 6.

Table 6
Coded Vectors for Linear-Model Analysis of Single-Test Validities

Cell Code*	TSSM				Item Type	Order	TSSM by Item Type				TSSM by Order				Type by Order	TSSM by Item Type by Order			
1	1	0	0	0	1	1	1	0	0	0	1	0	0	0	1	1	0	0	0
2	1	0	0	0	1	-1	1	0	0	0	-1	0	0	0	-1	-1	0	0	0
3	1	0	0	0	-1	1	-1	0	0	0	1	0	0	0	-1	-1	0	0	0
4	1	0	0	0	-1	-1	-1	0	0	0	-1	0	0	0	1	1	0	0	0
5	0	1	0	0	1	1	0	1	0	0	0	1	0	0	1	0	1	0	0
6	0	1	0	0	1	-1	0	1	0	0	0	-1	0	0	-1	0	-1	0	0
7	0	1	0	0	-1	1	0	-1	0	0	0	1	0	0	-1	0	-1	0	0
8	0	1	0	0	-1	-1	0	-1	0	0	0	-1	0	0	1	0	1	0	0
9	0	0	1	0	1	1	0	0	1	0	0	0	1	0	1	0	0	1	0
10	0	0	1	0	1	-1	0	0	1	0	0	0	-1	0	-1	0	0	-1	0
11	0	0	1	0	-1	1	0	0	-1	0	0	0	1	0	-1	0	0	-1	0
12	0	0	1	0	-1	-1	0	0	-1	0	0	0	-1	0	1	0	0	1	0
13	0	0	0	1	1	1	0	0	0	1	0	0	0	1	1	0	0	0	1
14	0	0	0	1	1	-1	0	0	0	1	0	0	0	-1	-1	0	0	0	-1
15	0	0	0	1	-1	1	0	0	0	-1	0	0	0	1	-1	0	0	0	-1
16	0	0	0	1	-1	-1	0	0	0	-1	0	0	0	-1	1	0	0	0	1
17	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1
18	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	1	1	1
19	-1	-1	-1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	-1	1	1	1	1
20	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	-1	-1	-1	-1

*See Figure 2 for definition of cell codes.

Linear contrasts. The elements of each coded vector in a linear-model analysis serve to define a linear contrast (Winer, 1971, p. 171) among the sample statistics associated with the various treatment combinations in the ANOVA layout. For example, in Table 6 the first of the 19 coded vectors generated for the linear-model analysis contains the following elements: 1, 1, 1, 1,...(12 zeroes)..., -1, -1, -1, -1. Since the elements of each coded vector were associated in subsequent steps of the analysis with the elements of the ordered vector $\underset{\sim}{z}$, the first coded vector implicitly defines a contrast between the $\underline{z}^*$ statistics associated with cells 1, 2, 3, and 4 (the BAYES tests) and the $\underline{z}^*$ statistics associated with cells 17, 18, 19, and 20 (the ASVAB/N tests).

The second, third, and fourth coded vectors served to contrast STMI, ASVAB/B, and ASVAB/M, respectively, with ASVAB/N in a similar manner. In each case, four $\underline{z}^*$ statistics were assigned a contrast coefficient of +1 and the $\underline{z}^*$ statistics of the ASVAB/N tests were assigned contrast coefficient values of -1. (All contrast coefficients in a coded vector were ultimately divided by the number of elements equal to +1 in that vector. This rescaled each contrast so that it reflected the difference between the mean $\underline{z}^*$ statistic for the set of cells originally weighted +1 and the mean $\underline{z}^*$ statistic for the set of cells originally weighted -1.)

The first four coded (contrast coefficient) vectors in Table 6 define contrasts that exhaust the four degrees of freedom available for an overall test of the main effect of TSSM. The fifth and sixth coded vectors contain contrast coefficients for contrasts that exhaust the two (single) degrees of freedom available for tests of the main effects of Item Type and Order. The fifth coded vector contains the elements 1, 1, -1, -1, ..., 1, 1, -1, -1, which contrasts AR tests with WK tests. The sixth vector contains the elements 1, -1, 1, -1, ..., 1, -1, 1, -1, which contrasts tests from the first half of the testing session with tests from the second half.

The coded vectors for interaction among the three independent variables were obtained by generating the element-by-element product of pairs of main-effect vectors. Thus, whenever either +1 or -1 appeared in the same position of two main-effect vectors for two different independent variables, a +1 appeared in the corresponding position of the resulting interaction vector. Whenever a +1 appeared in one main-effect vector and a -1 appeared in the same position of a main-effect vector for a different independent variable, a -1 appeared in the corresponding position of the resulting interaction vector. Each of the 13 interaction vectors in Table 6 defines a particular contrast among the $\underline{z}^*$ statistics associated with the cells of the 5 x 2 x 2 cross-classification. The complete set of 19 coded vectors in Table 6 defines a set of contrasts that exhaust the 19 degrees of freedom available in a 5 x 2 x 2 linear-model analysis of single-test $\underline{z}^*$ statistics.

Estimating the variance-covariance matrix. The asymptotic correlation (over random samples) between two $\underline{z}^*$ statistics calculated in a single sample from a trivariate-normal population with population correlations ρ_{12}, ρ_{13}, and ρ_{23} (where variable 1 is the criterion measure and variables 2 and 3 are predictor variables) is given by

$$\rho(z^*_{12}, z^*_{13}) = \frac{\rho_{23}(1 - \rho^2_{12} - \rho^2_{13}) - [\rho_{12}\rho_{13}(1 - \rho^2_{12} - \rho^2_{13} - \rho^2_{23})]/2}{(1 - \rho^2_{12})(1 - \rho^2_{13})} \qquad [27]$$

(Dunn and Clark, 1969). Thus, a large-sample estimate of the covariance between z^*_{12} and z^*_{13} values calculated in the same sample is provided by inserting large-sample estimates of ρ_{12}, ρ_{13}, and ρ_{23} on the right side of Equation 27 and multiplying the resulting value by 1/(N-3), the product of the asymptotic standard deviations of two *z** statistics calculated in a sample of size N.

Using the preceding asymptotic result, a large-sample estimate of the variance-covariance matrix among the *z** values in the vector $\underset{\sim}{z}$ was generated. First, each element of the matrix that represented a covariance between two *z** statistics computed in two independent samples rather than the same sample (i.e., the covariance between a *z** statistic from the WK-AR group and a *z** statistic from the AR-WK group) was set equal to zero. These are "known" covariance values. Then, the "unknown" within-group covariances were estimated using Equation 27 and the sampling variances on the diagonal were set equal to 1/(231-3) for *z** statistics computed in the WK-AR group and 1/(221-3) for *z** statistics computed in the AR-WK group. Thus, the 20 x 20 estimated variance-covariance matrix among the sample *z** statistics for single tests contained 20 known variance elements, 190 known between-group covariance elements, and 190 estimated within-group covariance elements.

If the estimated variance-covariance matrix among the elements of $\underset{\sim}{z}$ is designated as $\hat{\underset{\sim}{\Sigma}}_z$, a large-sample estimate of the variance-covariance matrix among a set of linear combinations of the elements of $\underset{\sim}{z}$ is given by

$$\hat{\underset{\sim}{\Sigma}}_c = \underset{\sim}{W}'\hat{\underset{\sim}{\Sigma}}_z\underset{\sim}{W} , \qquad [28]$$

where the columns of $\underset{\sim}{W}$ contain the fixed (over samples) coefficients for the linear combinations. Subsets of the 19 coded (contrast coefficient) vectors described previously were used to define various $\underset{\sim}{W}$ matrices. For example, in connection with the overall test of the main effect of TSSM, the first four vectors among the set of 19 coded vectors in Table 6 were rescaled (divided by 4) and the rescaled vectors were used to define a $\underset{\sim}{W}$ matrix with 20 rows and 4 columns. In this case, the resulting $\hat{\underset{\sim}{\Sigma}}_c$ was a large-sample estimate of the variance-covariance matrix, over random samples, among the four contrasts associated with TSSM. In a similar manner, $\hat{\underset{\sim}{\Sigma}}_c$ matrices were constructed to be used in connection with significance tests of the main effects of Item Type and Order and also for the various interactions of the three independent variables.

If the mean population *z** value (over levels of Item Type and Order) for BAYES is equal to the mean population *z** value for ASVAB/N, then the population value of the contrast defined by the first vector in Table 6 is zero. Similarly, if the mean population *z** value for STMI is equal to the mean population *z** value for ASVAB/N, the population value of the contrast defined by the second coded vector is zero. The same holds for the contrasts defined by the third and fourth coded vectors if

ASVAB/B and ASVAB/M have mean population $\underline{z}*$ values equal to that of ASVAB/N. Thus, a test of the hypothesis that the means (over levels of Item Type and Order) of the population $\underline{z}*$ values for all five TSSM combinations are equal is equivalent to a test of the hypothesis that the population values of the contrasts defined by the first four coded vectors in Table 5 are simultaneously equal to zero.

Test statistics. Sympson (in preparation) demonstrates that $\underline{p}$-element vectors of contrasts among sample $\underline{z}*$ statistics are asymptotically distributed $\underline{p}$-variate normal with expectation equal to the corresponding vector of contrasts among the population $\underline{z}*$ statistics (i.e., the sample vector is asymptotically unbiased as an estimate of the population vector). If the sample contrast vector $\underset{\sim}{W}'\underset{\sim}{z}$, which is 4 x 1 in the case of the four contrasts associated with TSSM, is distributed (approximately) $\underline{p}$-variate normal over random samples with expectation equal to the $\underline{p}$-element null vector and variance-covariance matrix $\underset{\sim}{\Sigma}_c$, the statistic

$$(\underset{\sim}{W}'\underset{\sim}{z})'\underset{\sim}{\Sigma}_c^{-1}(\underset{\sim}{W}'\underset{\sim}{z}) \qquad [29]$$

is distributed (approximately) as χ^2 with $\underline{p}$ degrees of freedom (Rao, 1973, p. 524). Now consider the statistic

$$S_1 = (\underset{\sim}{W}'\underset{\sim}{z})'\,\hat{\underset{\sim}{\Sigma}}_c^{-1}(\underset{\sim}{W}'\underset{\sim}{z}), \qquad [30]$$

where a consistent estimate of $\underset{\sim}{\Sigma}_c$ replaces the population variance-covariance matrix among the contrasts. Since $\hat{\underset{\sim}{\Sigma}}_c$ approaches $\underset{\sim}{\Sigma}_c$ as the number of observations associated with $\hat{\underset{\sim}{\Sigma}}_c$ increases, the asymptotic distribution of S_1 is also χ^2 with $\underline{p}$ degrees of freedom.

Sympson (in preparation) shows that for large samples the asymptotic sampling distribution of any monotone transformation of S_1 can be approximated using an identical transformation of Hotelling's T^2 statistic (Rao, 1973, p. 541). This fact suggests the possibility of using as a test statistic some transformation of S_1 that corresponds to a previously derived transformation of T^2. For example, in situations where all the elements of $\underset{\sim}{\Sigma}_z$ are unknown but are estimated using a single large sample, the test statistic

$$S_2 = [(N - p)/(N - 1)][S_1/p], \qquad [31]$$

where N is the number of individuals in the sample and $\underline{p}$ is the number of contrasts involved, can be assumed to follow (approximately) the F distribution with $\underline{p}$ and (N-$\underline{p}$) degrees of freedom (Morrison, 1967, p. 120).

The situation in this research is somewhat more complex in that $\hat{\underset{\sim}{\Sigma}}_z$ contained both known and unknown elements, and the unknown elements were estimated in two groups of slightly different size (N = 221 and N = 231). As a result, it is not obvious what should be the exact degrees of freedom associated with $\hat{\underset{\sim}{\Sigma}}_c$, whose elements are linear combinations of the 210 known and 190 unknown elements of $\hat{\underset{\sim}{\Sigma}}_z$. Since for large values of N the value of S_2 and the value of $F[\underline{p}, (N-\underline{p}), 1 - \alpha]$, where the latter quantity is the $100(1 - \alpha)\%$ point of an F distribution with $\underline{p}$ and (N - $\underline{p}$) degrees of freedom, are not much influenced by small changes in N, it was decided to proceed as though $\hat{\underset{\sim}{\Sigma}}_c$ had degrees of freedom equal to (221 + 231)/2. Thus, the statistic S_2 was treated as having $\underline{p}$ and (226 - $\underline{p}$) degrees of freedom. For each major hypothesis of interest, this approach provided a slightly more conservative test (i.e., lower probability of Type I error) than would be obtained by referring the statistic S_1 to the χ^2 distribution with $\underline{p}$ degrees of freedom.

Thus, to recapitulate, in the case of the significance test for the main effect of TSSM, the matrix $\underset{\sim}{W}$ was created from the first four vectors in Table 6. These vectors were rescaled by dividing the elements of each vector by the number of +1 values in the vector (i.e., 4). Then, the sample contrast vector $\underset{\sim}{W}'\underset{\sim}{z}$ and the estimated variance-covariance matrix $\hat{\underset{\sim}{\Sigma}}_c = \underset{\sim}{W}'\hat{\underset{\sim}{\Sigma}}_z\underset{\sim}{W}$ were computed. Finally, a value of the statistic S_1 was computed and entered into Equation 31 with N = 226 and $\underline{p} = 4$. It was assumed that the resulting value of S_2 would follow (approximately) the F distribution with 4 and 222 degrees of freedom if all four population contrasts were equal to zero. A computer program was used to determine the probability of observing an F statistic as large or larger than S_2 if the null hypothesis were true. If this probability was less than or equal to .05, the main effect for TSSM was declared statistically significant.

Similar procedures were followed for testing the main effect of Item Type and Order and for testing the various interactions among the three independent variables. For each such test, the matrix $\underset{\sim}{W}$ was defined using the appropriate set of (rescaled) vectors from Table 6 and the value of $\underline{p}$ in Equation 31 was set equal to the number of columns in $\underset{\sim}{W}$. If the probability of the obtained S_2 value was less than or equal to .05, the null hypothesis of no main effect or no interaction effect was rejected.

A posteriori significance tests. Following tests of the major hypotheses of interest (overall main effect and interaction tests), one-degree-of-freedom a posteriori significance tests were conducted. This was done by testing each of the individual contrasts originally associated with a significant major hypothesis, and also the set of contrasts obtained by taking pairwise differences among the original contrasts. For example, since the main effect for TSSM was found to be statistically significant, each of the four contrasts originally associated with the overall main-effect test was examined. For each of the original contrasts, a new matrix $\underset{\sim}{W}$ was defined that contained only the rescaled coefficients from that contrast's column in Table 6. Then, values of $\hat{\underset{\sim}{\Sigma}}_c$, S_1, and S_2 were computed using $\underset{\sim}{W}$, $\underset{\sim}{z}$, and $\hat{\underset{\sim}{\Sigma}}_z$. (Note that since $\underline{p} = 1$, $S_1 = S_2$ in this case.) Each of the contrasts defined by rescaling the coefficients in columns 1, 2, 3, and 4 of Table 6 was tested individually in this manner.

Next, six new vectors of contrast coefficients were generated by taking pairwise differences among the first four vectors in Table 6. For example, the pairwise difference between the first and second columns of Table 6 generated a contrast coefficient vector with the elements 1, 1, 1, 1, -1, -1, -1, -1, ... (12 zeroes). This coefficient vector, after rescaling, was used to define a new $\underset{\sim}{W}$ and a test was conducted of the hypothesis that the mean population z^* value for BAYES was equal to the mean population z^* value for STMI. Similar tests were conducted for the five remaining contrasts obtained by computing pairwise differences among the original contrast coefficient vectors. Testing the four contrasts defined by the first four rescaled columns of Table 6 and the six contrasts obtained by taking pairwise differences among these four columns gave a total of 10 single-degree-of-freedom a posteriori tests that were conducted following the significant overall test of the main effect of TSSM. These a posteriori tests systematically contrasted all possible pairings of the five TSSM combinations.

A similar a posteriori test procedure was followed after obtaining a significant result in the overall 3-way interaction test. In this case, each of the contrasts defined by the last four (rescaled) columns in Table 6 was tested individually, as were the six contrasts obtained by taking pairwise differences among these columns.

Whenever a set of a posteriori tests was executed following a significant major hypothesis test, the individual contrasts were not declared to be significantly different from zero unless the probability of the observed value of S_2 for that contrast was less than or equal to .05 divided by the number of a posteriori tests in the set. Thus, in order for one of the individual TSSM main-effect contrasts or one of the three-way interaction contrasts to be declared significant, its probability under the null hypothesis had to be less than or equal to .05/10 = .005.

The technique of testing individual contrasts at reduced Type I error (alpha) levels that was used here has been referred to in the statistical literature as Dunn's procedure and/or the Bonferroni test procedure (e.g., Kirk, 1968, p. 79) and, in the case of pairwise contrasts among marginal means in an ANOVA, Fisher's modified least-significant-difference (modified LSD) approach (e.g., Winer, 1971, p. 199). The major advantages of this approach are that the contrasts tested need not be orthogonal, the statistics involved need not be based on equal sample sizes, and the Type I error rate for the set of contrasts is less than or equal to the overall alpha level set by the experimenter (e.g., .05).

Validities for Composites

The statistical inference procedures described previously in connection with the analysis of single-test validities were also used in the analysis of composite (AR+WK) test validities. Two types of composites were studied: fixed-weight composites and optimally-weighted composites. Fixed-weight-composite scores were obtained by computing the mean of the AR and WK test scores obtained by an individual under each TSSM. Optimally-weighted-composite scores were obtained by computing under each TSSM the least-squares multiple linear regression estimate (Draper & Smith, 1966) of each individual's criterion score using the individual's AR and WK scores as predictors.

The computation of composite scores collapsed AR and WK scores into a single variable; therefore, the three-way cross-classification shown in Figure 2 became a two-way cross-classification. Moreover, since each composite score was based on one test that was administered before the 5-minute break and one test that was administered after the break, the earlier concept of order of administration (first half of testing session versus second half) was no longer applicable. Instead, the second independent variable in the analysis of composite validities was Content Order (AR test before WK test or vice-versa). The cell codes assigned in the two-way cross-classification of composite scores are shown in Table 7. Associated contrast coefficient vectors appear in Table 8.

Table 7
Cell Codes in Two-Way
Cross-Classification for
Composite-Score Validities

	Content Order	
TSSM	AR-WK	WK-AR
BAYES	1	2
STMI	3	4
ASVAB/B	5	6
ASVAB/M	7	8
ASVAB/N	9	10

As mentioned earlier, Content Order in the two-way cross-classification is logically related to the Item Type by Order two-way interaction in the three-way cross-classification for single-test validities. This relationship becomes apparent from examination of the fifteenth coded vector in Table 6, which shows that the contrast defined by this vector adds $\underline{z}^*$ statistics obtained from AR tests given before the 5-minute break together with $\underline{z}^*$ statistics obtained from WK tests given after the break, then contrasts the resulting sum with the sum of $\underline{z}^*$ statistics obtained from WK tests given before the break and AR tests given after the break. The sign of the observed sample contrast indicates which content order (AR-WK or WK-AR) resulted in the largest mean $\underline{z}^*$ value for single tests.

The Content Order main-effect contrast for the two-way analyses (the fifth coded vector in Table 8) generates a similar statistic based on composite-score validities. Note that the fifteenth coded vector in Table 6 and the fifth coded vector in Table 8 define the only between-groups contrasts in these two tables. As mentioned earlier, hypothesis tests associated with these contrasts will usually be less sensitive to the presence of real population effects than tests that are based on within-group contrasts. Note also that as a result of the relationship between the Content Order main effect in the two-way analyses and the Item Type by Order two-way interaction in the three-way analysis, a relationship also exists between the TSSM by Content Order two-way interaction in the two-way analyses and the TSSM by Item Type by Order three-way interaction in the analysis of single-test validities.

Table 8
Coded Vectors for Linear-Model Analyses of Composite-Score Validities

Cell Code*	TSSM				Content Order	TSSM by Content Order			
1	1	0	0	0	1	1	0	0	0
2	1	0	0	0	-1	-1	0	0	0
3	0	1	0	0	1	0	1	0	0
4	0	1	0	0	-1	0	-1	0	0
5	0	0	1	0	1	0	0	1	0
6	0	0	1	0	-1	0	0	-1	0
7	0	0	0	1	1	0	0	0	1
8	0	0	0	1	-1	0	0	0	-1
9	-1	-1	-1	-1	1	-1	-1	-1	-1
10	-1	-1	-1	-1	-1	1	1	1	1

*See Table 7 for definition of cell codes.

Fixed-weight composites. The linear-model analysis for fixed-weight composites was conducted in exactly the same manner as the analysis for single-test validities. Since fixed weights do not vary over samples, a composite formed using prespecified weights can be treated as a new predictor variable, and any validity coefficient obtained for the composite can be treated as though it were a "single-test" validity. Thus, after computing the sample correlation matrix among fixed-weight-composite scores and the validities of these scores with respect to the end-of-course criterion, large-sample estimates of the correlations (over samples) among z^* transformations of the validities were generated using Equation 27. These correlations were then transformed to covariances by multiplying by 1/(N-3) within each group (the AR-WK and WK-AR groups) and the estimated within-group covariances were used, along with the known between-group covariances of zero and the known sampling variances of 1/(N-3), to construct a large-sample estimate of the variance-covariance matrix among z^* statistics derived from the fixed-weight composites.

Given the matrix $\hat{\Sigma}_z$, hypothesis tests regarding the effects of TSSM, Content Order, and their two-way interaction were conducted by selecting appropriate sets of coded vectors from Table 8, rescaling the vectors, forming different $\underset{\sim}{W}$ matrices depending on the hypothesis being tested, computing Equation 28 for each hypothesis test, and then computing the values of S_1 and S_2 (Equations 30 and 31) for each hypothesis test. If the probability of an observed S_2 value was less than or equal to .05 under the associated null hypothesis, that hypothesis was rejected. After observing a significant S_2 value, single-degree-of-freedom contrasts were generated and tested in the manner that was described earlier in connection with the analysis of single-test validities.

Optimally-weighted composites. The linear-model analysis for optimally-weighted composites proceeded in a somewhat different fashion than that for fixed-weight composites. Since the weights obtained in least-squares prediction vary over random samples from a population of individuals, optimally-weighted-composite scores obtained in any one sample cannot be treated as if they were predictor scores drawn from a bivariate population of predictor and criterion scores. Individuals' optimally-weighted-composite scores depend not only on their test performance but also on the particular random sample of individuals in which they happen to be imbedded. If the same individual were observed as a member of a different sample, the person's composite score would change due to the use of new "optimal" weights derived in the second sample. This may be contrasted with the situation where fixed-weight-composite scores are generated. In that case, the combining weights do not vary over samples and an individual's composite score remains the same regardless of the sample in which the person is imbedded. Thus, in the case of a fixed-weight composite, an observed sample of individuals can be treated as a random sample from a bivariate population of predictor and criterion scores.

The objective of the analysis of optimally-weighted-composite scores was to determine whether TSSM, Content Order, and/or their interaction had a significant effect on the level of predictive validity of these composites. The population validity of an optimally-weighted composite is indexed by the population multiple correlation R. Consider the population contrast

$$\Psi = \sum_{i=1}^{k} w_i R_i^2 \, , \qquad [32]$$

where R_i^2 is the population squared multiple correlation obtained under the $\underline{i}$th "condition" (e.g., a particular TSSM in conjunction with a particular Content Order), w_i is a contrast coefficient specified for condition $\underline{i}$, and

$$\sum_{i=1}^{k} w_i = 0 \, . \qquad [33]$$

Sympson (in preparation) shows that whenever individuals are assigned to between-group conditions at random and then observed under each possible within-group condition, the population contrast

$$\Gamma = \sum_{i=1}^{k} w_i[\sigma^2(y - \hat{y}_i)] \qquad [34]$$

is equal to zero whenever the corresponding contrast Ψ is equal to zero. In Equation 34, $\sigma^2(\underline{y} - \hat{y}_i)$ is the population residual variance in $\underline{y}$, the criterion variable, given that $\hat{y}_i$, the population least-squares regression estimate of $\underline{y}$ under condition $\underline{i}$, has been partialled from $\underline{y}$. Thus, in order to test the hypothesis $\Psi = 0$ for any Ψ of interest, it is sufficient to test the hypothesis $\Gamma = 0$ which corresponds to that Ψ. Sympson (in preparation) also shows that an unbiased estimate of Γ is given by

$$\hat{\Gamma} = \sum_{i=1}^{k} w_i \mathrm{MSE}_i = \sum_{i=1}^{k} w_i \left(\frac{\mathrm{SSE}_i}{N_i - p_i - 1} \right), \qquad [35]$$

where MSE_i is the unbiased sample mean-square-error of estimate under condition $\underline{i}$ and p_i is the number of predictors under condition $\underline{i}$.

Sympson (1979) has indicated that a large-sample estimate of the sampling variance of an MSE is provided by

$$2[\mathrm{MSE}_i]^2/(N_i - p_i - 1) \qquad [36]$$

and that a large-sample estimate of the sampling covariance of two MSEs computed in the same sample is provided by

$$r(e_i^2, e_j^2)\left(\frac{2[\mathrm{MSE}_i][\mathrm{MSE}_j]}{[(N_i - p_i - 1)(N_j - p_j - 1)]^{\frac{1}{2}}} \right). \qquad [37]$$

In Equation 37, $r(e_i^2, e_j^2)$ is the sample product-moment correlation between the squared errors of estimate observed under condition $\underline{i}$ and the squared errors of estimate observed under condition $\underline{j}$. Of course, if two MSEs are computed in independent samples, and not in the same sample, their sampling covariance is known to be zero and need not be estimated. (Sympson, in preparation, derives a better asymptotic estimate of the covariance between two MSEs, but Equation 37 was used here.)

Using Equations 36 and 37, a large-sample estimate of the variance-covariance matrix among the sample MSEs was constructed. There were $\underline{k} = 10$ MSEs of interest, one for each cell in the two-way classification shown in Table 7. Thus, a 10 x 10 estimated variance-covariance matrix, $\hat{\underset{\sim}{\Sigma}}_M$, containing 50 known between-group covariance elements (zeroes), 40 estimated within-group covariance elements, and 10 estimated sampling variances was constructed. The sample MSE values themselves were arrayed in a 10-element vector, $\underset{\sim}{m}$, in the order of the cell codes shown in Table 7.

Sympson (in preparation) shows that $\underline{k}$-element vectors of sample MSE values are asymptotically distributed $\underline{k}$-variate normal with expectation equal to the corresponding vector of population residual variances. Thus, $\underline{p}$-element sample contrast vectors of the type

$$\hat{\underset{\sim}{\Gamma}} = \underset{\sim}{W}'\underset{\sim}{m} , \qquad [38]$$

where $\underset{\sim}{W}$ contains a $\underline{k}$-element vector of contrast coefficients in each of its $\underline{p}$ columns, will be asymptotically distributed $\underline{p}$-variate normal with expectation equal to the corresponding population contrast vector (Bock, 1975, p. 141). Moreover, assuming $\hat{\underset{\sim}{\Sigma}}_M$ is a consistent estimate of the variance-covariance matrix $\underset{\sim}{\Sigma}_M$, the matrix

$$\hat{\underset{\sim}{\Sigma}}_c = \underset{\sim}{W}'\hat{\underset{\sim}{\Sigma}}_M\underset{\sim}{W} \qquad [39]$$

provides a consistent estimate of the variance-covariance matrix among the sample contrasts.

The preceding results allow use of the statistics S_1 and S_2 described earlier (Equations 30 and 31) to test joint null hypotheses for sets of contrasts among residual variances. In computing the sample statistic S_1 (Equation 30), $\hat{\underset{\sim}{\Sigma}}_c$ is as defined above and the vector $\underset{\sim}{m}$ replaces the vector $\underset{\sim}{z}$. In computing S_2 (Equation 31), a value for N, the degrees of freedom associated with $\hat{\underset{\sim}{\Sigma}}_c$, must be specified. $\underset{\sim}{W}$ was defined for each major hypothesis of interest by selecting the appropriate set of $\underline{p}$ contrast-coefficient vectors from Table 8 and rescaling them as before. The value of N was set equal to 223 and the hypothesis that the $\underline{p}$ population contrasts were simultaneously equal to zero was rejected if the obtained value of S_2 exceeded the 95th percentile of the F distribution with $\underline{p}$ and (223-$\underline{p}$) degrees of freedom. The main effects of TSSM and Content Order, and the interaction of these two independent variables, were each tested in this fashion.

In the linear-model analysis for optimally-weighted composites, 223 degrees of freedom were assumed for the matrix $\hat{\underset{\sim}{\Sigma}}_c$ since each MSE value was the mean of either 231 or 221 squared residuals (hence, approximately 226) and three parameters (an intercept and two slope coefficients) were estimated in order to obtain each MSE value. As noted earlier, the computed value of S_2 and the indicated critical value of S_2 will change only slightly as a result of small changes in N when the value of N is this large. Hence, the particular value assumed for the degrees-of-freedom parameter was not deemed a crucial consideration in this application.

Pre-enlistment ASVAB composites. As noted previously, pre-enlistment scores on five Air Force ASVAB composites (Mechanical, Administrative, General-Technical, Electronics, and AFQT) were available for 406 of the individuals in Subgroup 4 (see Table 4). For each of these ASVAB composites, criterion-related validity correlations were computed in the WK-AR group (N = 206), the AR-WK group (N = 200), and the Total group (N = 406) making up Subgroup 5. Validities for the experimental TSSM composites were also calculated in these three groups.

Effect of Limiting Boundary Values on the Validity of Maximum Likelihood Ability Estimates

Before proceeding with the various linear model analyses in this research, another issue had to be addressed. Since choice of a procedure for dealing with "out-of-bounds" maximum likelihood ability estimates can influence the correlation of such estimates with other variables, it was important to explore the effect on test validity of various methods for dealing with out-of-bounds cases. To this end, TSSM validity coefficients were computed under three different conditions. First, separate validities were calculated for the WK-AR and AR-WK groups in Subgroup 4 using the original limiting boundary values (±5.00) that had been imposed during the data collection process. Within each group, validity coefficients were computed for STMI AR at 20 items, STMI WK at 30 items, ASVAB/M AR, ASVAB/M WK, the fixed-weight STMI composite (AR+WK), and the fixed-weight ASVAB/M composite (AR+WK). Thus, 12 validity coefficients were computed using the original limiting boundaries.

Next, the same 12 validity coefficients were computed using only those members of Subgroup 4 who obtained "in-bounds" maximum likelihood ability estimates in STMI AR (at 20 items), STMI WK (at 30 items), ASVAB/M AR, and ASVAB/M WK. The resulting validity coefficients were based on 201 cases in the AR-WK group and 208 cases in the WK-AR group. Finally, the same 12 validities were computed with all out-of-bounds estimates set equal to new limiting values that were selected to be approximately .30 ability unit more extreme (more deviant from the mean ability estimate) than the most extreme in-bounds estimates obtained on a given test. The revised boundaries were -4.10 and 2.70 for STMI AR, -2.80 and 2.90 for STMI WK, -5.20 and 2.40 for ASVAB/M AR, and -4.40 and 3.10 for ASVAB/M WK. The validities computed using these boundaries were based on the same number of cases as the validities computed with the original boundaries (231 cases in the WK-AR group and 221 cases in the AR-WK group).

The revised boundaries just described are quasi-empirical-Bayes ability estimates in the sense that they were selected to be "not too distant" from the most extreme in-bounds estimates observed in the experimental samples. Another possible approach would be to determine a priori the lowest and highest finite $\hat{\theta}$ values that could be obtained from a particular test and then set boundary values "not too distant" from these finite extremes. To the extent that the most extreme in-bounds estimates obtained in an empirical sample approximate the most extreme finite $\hat{\theta}$ values obtainable from a particular test, the quasi-empirical-Bayes method for selecting boundaries used here and the a priori method based on maximally deviant finite $\hat{\theta}$ values will be equivalent. Another method for dealing with response vectors that generate infinite $\hat{\theta}$ values has been proposed by Samejima (1980, pp. 83-97).

Intercorrelations Between AR and WK Tests

During the data analysis, it was observed that product-moment correlations between AR and WK single-test scores were lower for the two adaptive testing strategies than for the ASVAB TSSMs. The statistical significance of differences among these correlations was examined using an asymptotic test statistic described by Dunn and Clark (their so-called "Best" test; 1969). For each difference tested, a large-sample estimate of the correlation (over random samples) between the correlation coefficients contrasted was computed using Dunn and Clark's Equation 9. The resulting large-sample estimates were then used in computing the asymptotic test statistics. This set of significance tests was based on all 452 individuals in Subgroup 4.

RESULTS

Preliminary Results

Fill-In Analysis

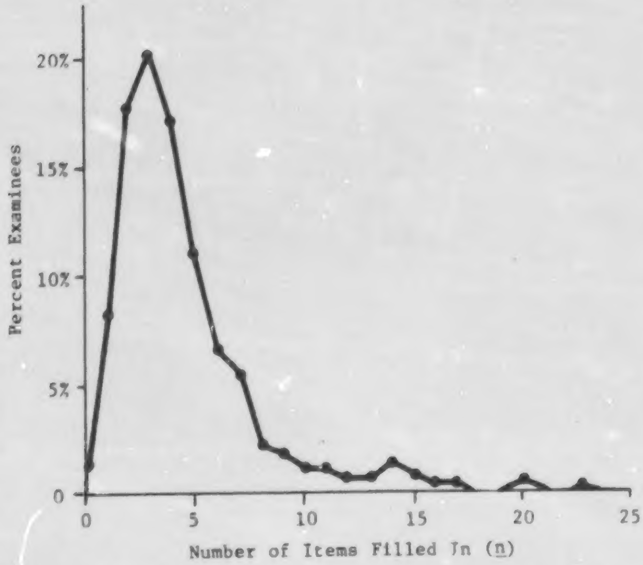
Figure 3 shows distributions of the number of items filled in during an examinee's second adaptive AR test (Figure 3a) and second adaptive WK test (Figure 3b). Each of these distributions is based on the 452 examinees in Subgroup 4. For the AR test, the minimum number of items filled in was 0, the maximum was 23, the mean was 4.39, and the mode was 3 out of the 25 items selected. For the WK test, the minimum number of items filled in was again 0, the maximum was 31, the mean was 4.50 and the mode was 4 items out of the 35 items selected. On the average, about 85% of the items selected in each second adaptive test were identical with those administered in the first test. These data indicate, therefore, that the two adaptive testing strategies--BAYES and STMI--generally selected the same items.

Nonconverged and Out-of-Bounds Maximum Likelihood Estimates

Figure 4 shows the percent of cases in Subgroup 4 with nonconverged ability estimates at each stage in the two STMI tests. Each test shows a high percentage of nonconvergences during the early stages of the test due to the constraints imposed on the number of maximum likelihood iterations. In both tests, the percentage of nonconvergences dropped rapidly to less than 5% of the examinees for tests of nine items or more. For the AR test, there were one or two nonconvergences for tests of 20 and 25 items in length, but for the WK test there were no nonconvergences for test lengths of more than 15 items.

Figure 3
Percent of Examinees Having n Items Filled In During
the Second Adaptive Test (N=452)

(a) AR



(b) WK

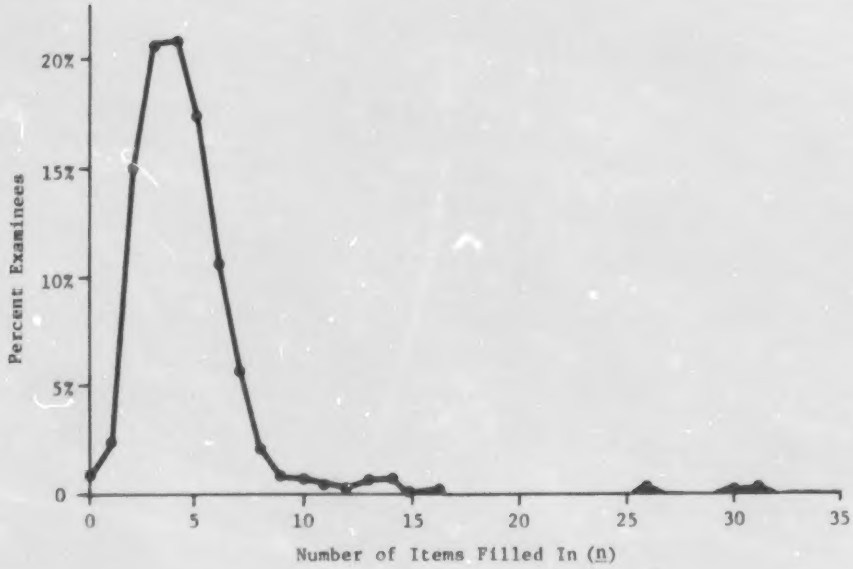
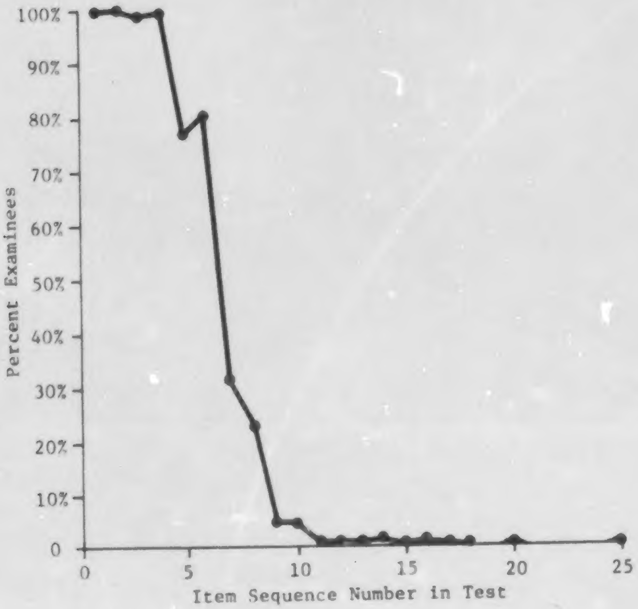


Figure 4
Percent of Examinees with Nonconverged Ability Estimates
at Each Stage in the STMI Tests (N=452)

(a) AR



(b) WK

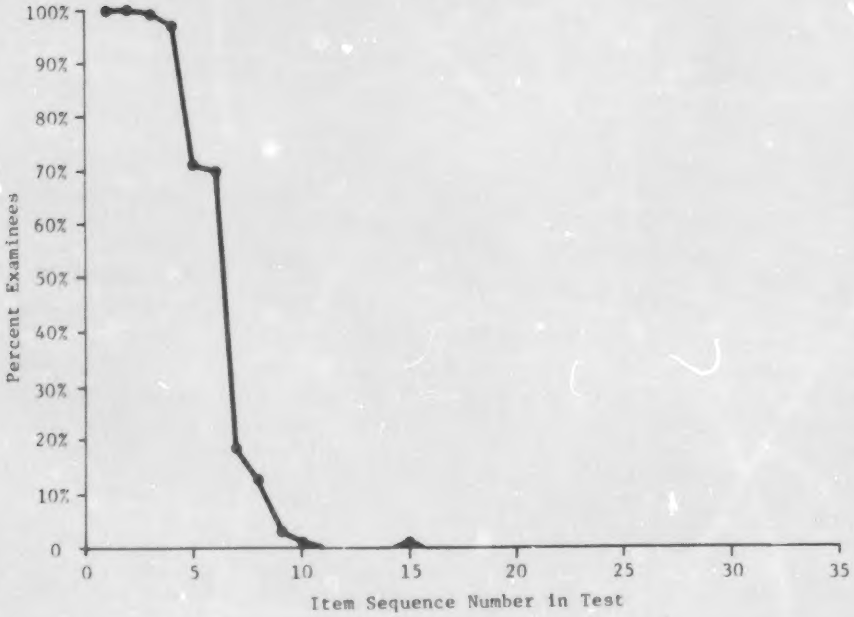


Figure 5 shows the number of out-of-bounds cases in Subgroup 4 at each stage of the STMI AR test. All such cases were at the lower boundary of -5. There were no out-of-bounds estimates at any test length in STMI WK. The figure shows an initial increase at item 3. Before this point in the test, the constraints imposed on the maximum likelihood iterations prevented any ability estimates from reaching either boundary. After the third item, typically only one or two cases among the 452 in this subgroup were out of bounds.

Figure 5
Number of Examinees with $\hat{\theta}$ of -5.0 at Each Stage of the STMI AR Test (Total N=452)

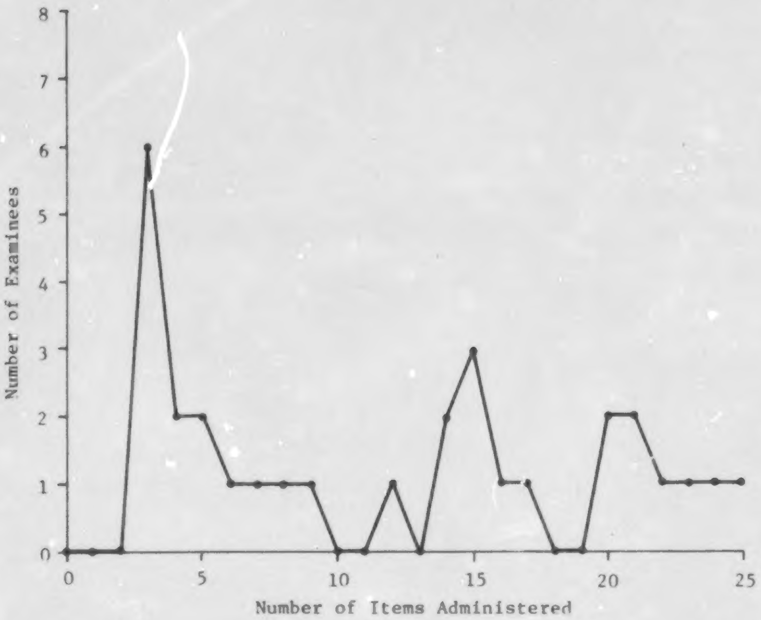


Table 9 shows the relative frequency of high and low out-of-bounds maximum likelihood ability estimates obtained in Subgroups 1 and 2 for the STMI AR test at 20 items, the STMI WK test at 30 items, and the ASVAB/M tests. As the data in Table 9 show, the relative frequency of out-of-bounds cases was 0 for the STMI WK test and was less than 1% in one instance for the STMI AR test. For the ASVAB/M AR test, the percentage of out-of-bounds cases varied from 2.9% to 4.9%. For the ASVAB/M WK test the percentage of out-of-bounds cases varied from less than 1% to 2.8%. Thus, for the adaptive tests, there were very few out-of-bounds cases in comparison to the conventional ASVAB tests scored by maximum likelihood.

Effect of Variable Entry

Figure 6 shows the correlations of ability estimates calculated after each item was administered with final ability estimates, separate-

Table 9

Relative Frequency in WK-AR and AR-WK Groups of Out-of-Bounds Low ($\hat{\theta} = -5.0$) and High ($\hat{\theta} = 5.0$) Maximum Likelihood Ability Estimates

		AR		WK	
Test	Group	Low	High	Low	High
STMI	WK-AR	2/245	0/245	0/247	0/247
	AR-WK	0/243	0/243	0/245	0/245
ASVAB/M	WK-AR	10/245	7/245	1/247	7/247
	AR-WK	8/243	12/243	1/245	3/245

ly for the BAYES and STMI AR and WK tests. As the data show, for BAYES (Figure 6a and 6b), and STMI WK (Figure 6d) the correlations of interim ability estimates with final ability estimates were similar for the fixed-entry and variable-entry conditions. In both cases, there was a tendency for the fixed-entry condition to have slightly lower correlations with final score through test lengths of about 15 to 20 items, but the differences were not very large. A different pattern emerged for the STMI AR test (Figure 6c). Contrary to the results for the other three tests, the correlations of interim ability estimates with final ability estimates were higher for the fixed-entry condition than for the variable-entry condition, and the differences were substantially larger.

Examinee Response Time

Table 10 gives summary statistics for MERT on the ASVAB, BAYES, and STMI tests in Subgroups 1 and 2. Figure 7 shows MERT relative frequency distributions for the three AR tests, and Figure 8 shows relative frequency distributions for the WK tests. The MERT distributions for the two adaptive tests were quite similar, since about 85% of the items selected were common to the two tests. As the data show, MERT for the ASVAB tests was shorter, on the average, than was MERT for the adaptive tests. For ASVAB AR, mean MERT was about 50 seconds, while the mean MERT for BAYES and STMI was about 68 seconds. Mean MERT for the ASVAB WK tests was about 10 seconds, whereas for the two adaptive WK tests mean MERT was about 14 seconds per item. Mean MERT was consistently 2 to 3 seconds less after the 5-minute break than before the break.

Computer Response Time

Table 11 provides summary statistics for MCRT for ASVAB, and Table 12 provides similar data for the two adaptive tests. Frequency distributions of MCRT for the three testing strategies are presented in Figure 9, separately for the AR and WK tests. As Table 11 shows, MCRT for the ASVAB had a mode of less than one second per item, with most of the MCRTs less than 1.5 seconds for WK tests (Figure 9c), and less than about 3 seconds for the AR test (Figure 9a). By contrast, the modal MCRT for the BAYES adaptive test (Table 12) was about 3.3 seconds for

Figure 6
Correlation of Interim Ability Estimates with Final Ability Estimates in the Variable-Entry and Fixed-Entry Conditions, for BAYES and STMI Tests

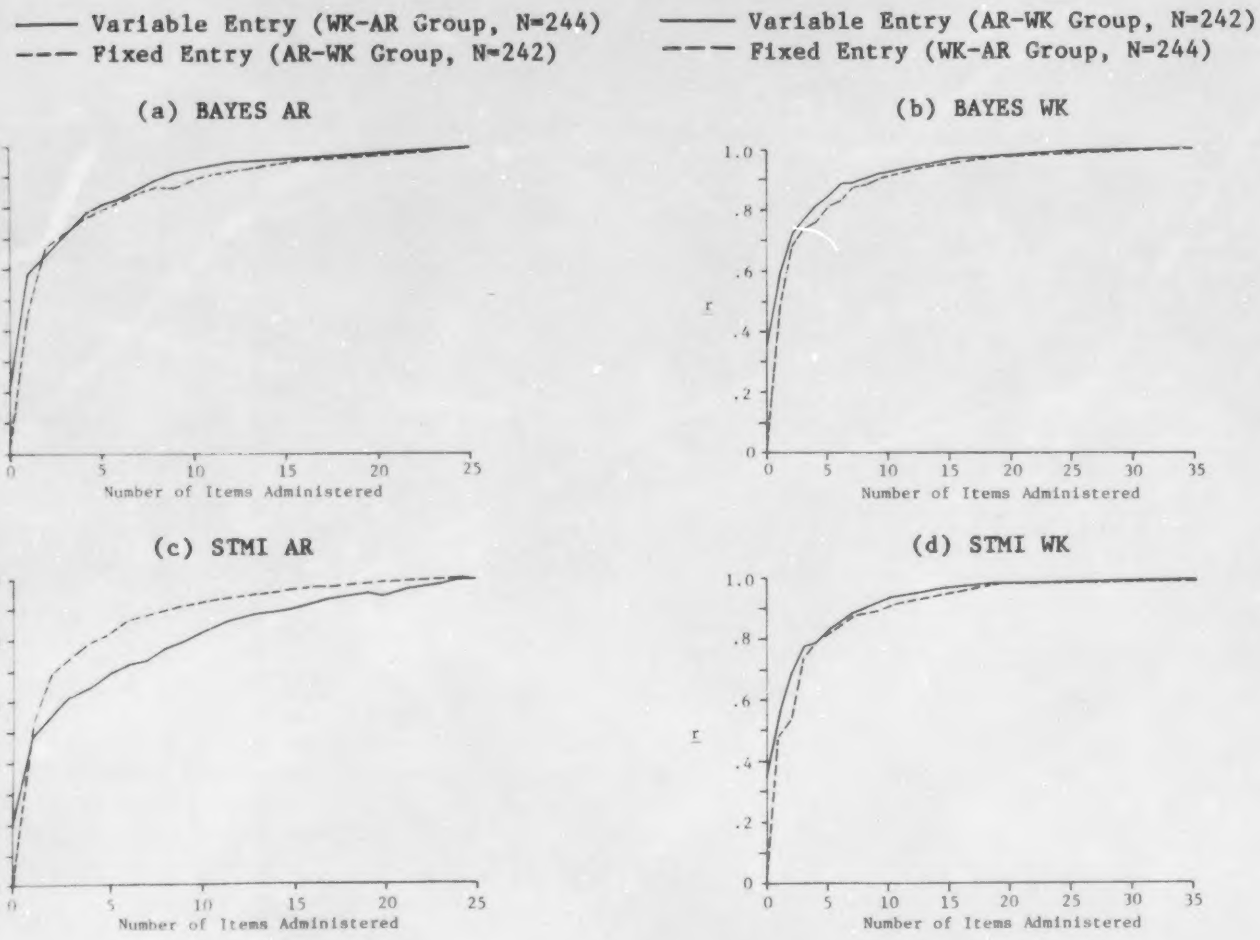


Table 10
Summary Statistics for Mean Examinee Response Time Distributions for ASVAB, BAYES, and STMI Tests with AR and WK Items, for WK-AR and AR-WK Groups

TSSM and Statistic	AR (Subgroup 1) WK-AR (N=245)	 AR-WK (N=243)	WK (Subgroup 2) AR-WK (N=245)	 WK-AR (N=247)
ASVAB				
Mean	48.739	51.004	9.734	11.676
S. D.	21.168	18.780	3.942	5.249
Skew	1.017	.977	2.445	1.742
Kurtosis	1.652	1.100	12.087	5.168
Minimum	10.200	12.410	3.533	4.087
Maximum	135.350	115.520	37.683	38.733
BAYES				
Mean	66.996	69.475	12.936	15.016
S. D.	29.219	28.847	4.247	5.221
Skew	1.383	1.111	1.795	1.501
Kurtosis	4.036	1.849	7.785	3.609
Minimum	17.156	19.472	5.746	6.054
Maximum	224.796	194.848	41.283	40.351
STMI				
Mean	66.863	70.094	13.062	15.271
S. D.	29.787	28.904	4.319	5.373
Skew	1.689	1.003	1.698	1.503
Kurtosis	6.158	1.448	6.700	3.626
Minimum	13.384	19.472	5.886	6.360
Maximum	246.724	192.916	40.851	40.814

Table 11
Summary Statistics for Mean Computer Response Time Distributions for ASVAB with AR and WK Items, for Adaptive-Test Order Groups

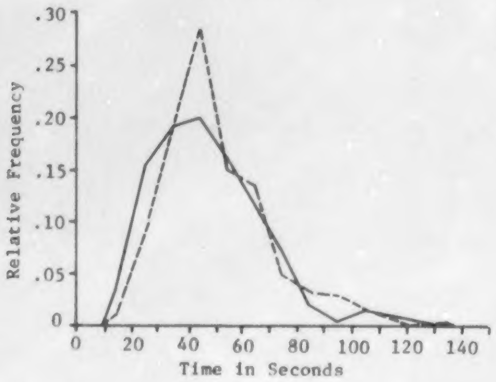
Statistic	AR (Subgroup 1) Group 1	 Group 2	WK (Subgroup 2) Group 1	 Group 2
N	244	244	246	246
Mode	.897	.890	.826	.830
Minimum	.775	.745	.633	.627
25%	.805	.791	.682	.676
50%	1.333	1.158	.864	.852
75%	3.160	2.569	1.354	1.126
90%	6.269	4.706	2.851	1.847

Note. Group 1, BAYES administered before STMI;
Group 2, STMI administered before BAYES.

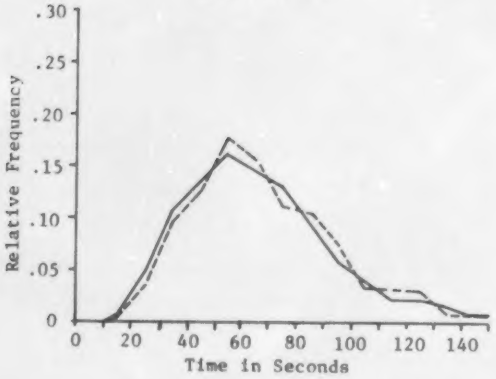
Figure 7
Frequency Distributions of Mean Examinee Response Time per Item
in ASVAB, BAYES, and STMI AR Tests

—— WK-AR Group (N=245) - - - AR-WK Group (N=243)

(a) ASVAB



(b) BAYES



(c) STMI

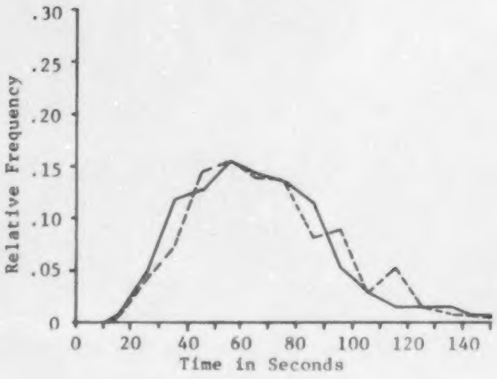
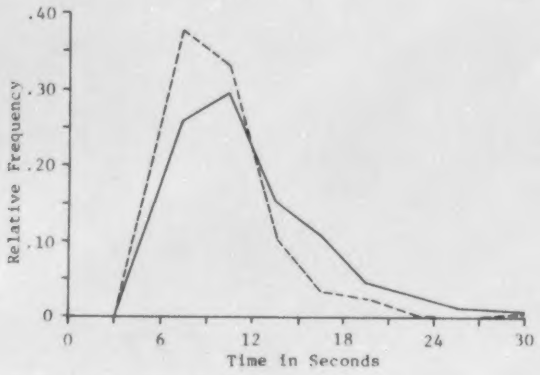


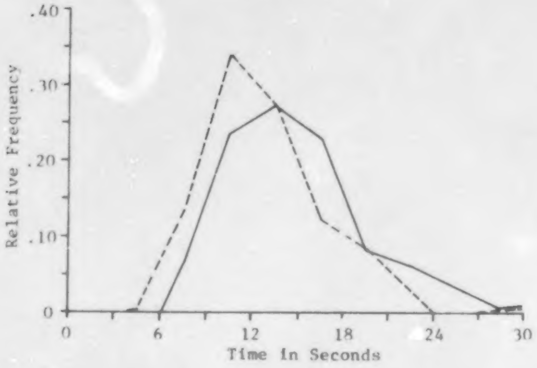
Figure 8
Frequency Distributions of Mean Examinee Response Time per Item in ASVAB, BAYES, and STMI WK Tests

——— WK-AR Group (N=247)   ---- AR-WK Group (N=245)

(a) ASVAB



(b) Bayes



(c) STMI

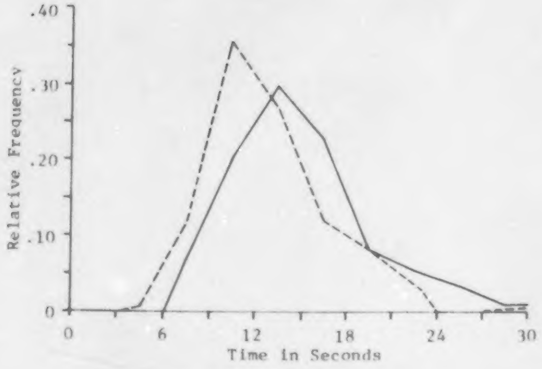
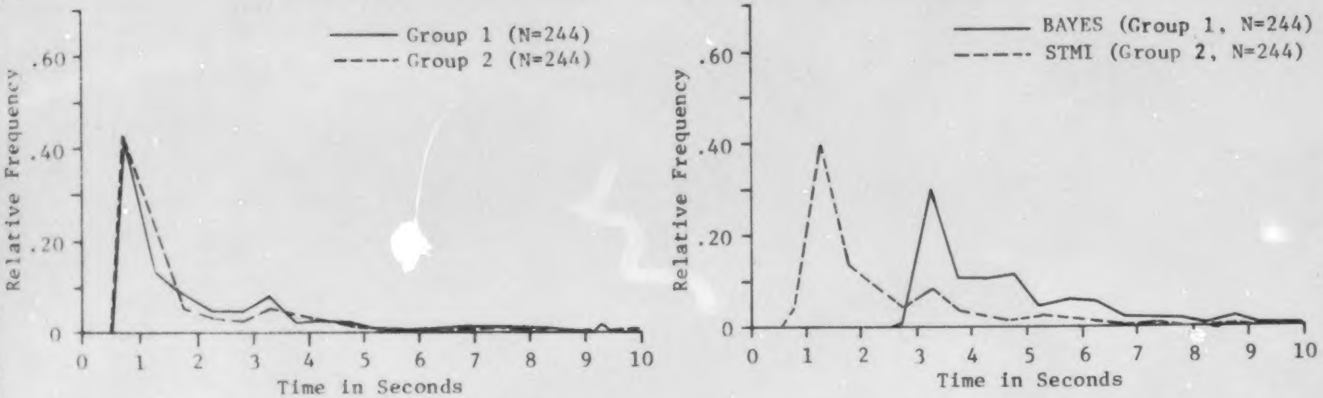
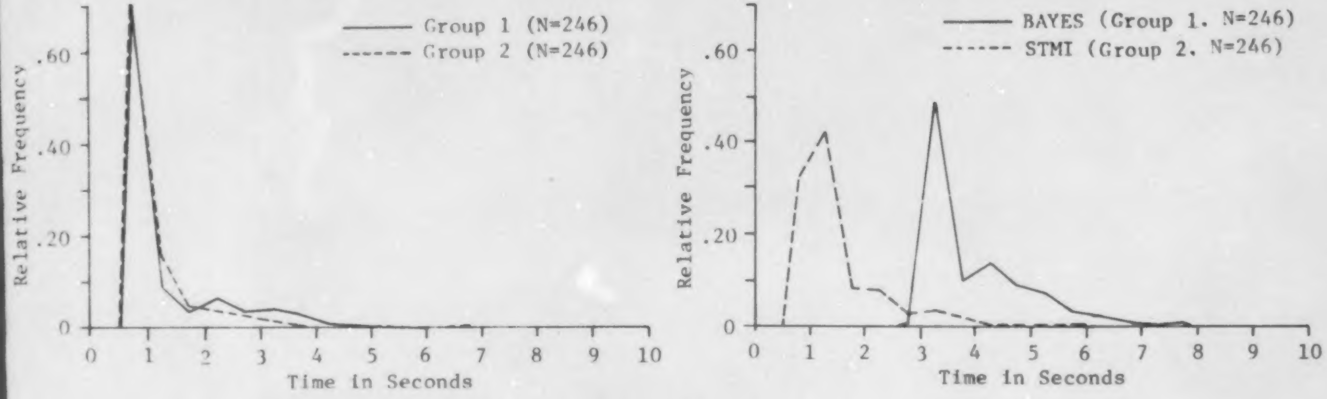


Figure 9
Frequency Distributions of Mean Computer Response Time per Item in ASVAB, BAYES, and STMI AR and WK Tests

(a) ASVAB AR **(b) BAYES AND STMI AR**



(c) ASVAB WK **(d) BAYES AND STMI WK**



both AR and WK, with 50% of the cases requiring at least 3.5 seconds of computer time for a WK response and 4.4 seconds for an AR response (Figures 9d and 9b). The results for the STMI tests were much more similar to that of the ASVAB tests, with modes of slightly more than 1 second. As Figure 9 shows, the distributions of MCRT for STMI had substantial overlap with those for the ASVAB test, while the MCRT for BAYES had very little overlap with that of the ASVAB or STMI tests.

Table 12
Summary Statistics for Mean Computer Response Time
Distributions for BAYES and STMI Adaptive Tests
with AR and WK Items

Statistic	AR (Subgroup 1) BAYES[a]	AR (Subgroup 1) STMI[b]	WK (Subgroup 2) BAYES[a]	WK (Subgroup 2) STMI[b]
N	244	244	246	246
Mode	3.299	1.287	3.276	1.119
Minimum	2.936	.964	2.894	.866
25%	3.390	1.260	3.245	.890
50%	4.364	1.711	3.499	1.210
75%	5.999	3.215	4.556	1.618
90%	8.895	5.300	5.455	2.409

[a] BAYES administered before STMI.
[b] STMI administered before BAYES.

Characteristics of Test Scores

Distributions

ASVAB number correct. Table 13 shows summary statistics for ASVAB number-correct score distributions for the AR and WK subtests in Subgroups 1 and 2; Figure 10 shows the frequency distributions for these subtests. As the data show, neither test was difficult for the group tested. For the AR test, the mean score was about 13, with a substantial portion of the examinees obtaining scores of 15 or more on the 20-item test (Figure 10a). The ASVAB WK test was even easier. On this test, the mean score was over 22, with a large portion of the examinees obtaining scores between 22 and the maximum score of 30 (Figure 10b).

IRT ability estimates. Table 13 also provides summary statistics for the IRT ability estimates derived from rescoring of the ASVAB by Bayesian and maximum likelihood methods (ASVAB/B and ASVAB/M), and for the BAYES and STMI adaptive tests. Frequency distributions for the AR tests are in Figure 11, and for the WK tests in Figure 12. Cases with out-of-bounds maximum likelihood estimates are not shown in these figures. For the AR data, mean ability estimates were about -.7 for ASVAB/B and -.4 for ASVAB/M. BAYES resulted in a mean ability estimate for AR of about -.5 and STMI about -.4. Thus, the mean ability levels resulting from all four tests were similar and the data indicate that

Table 13

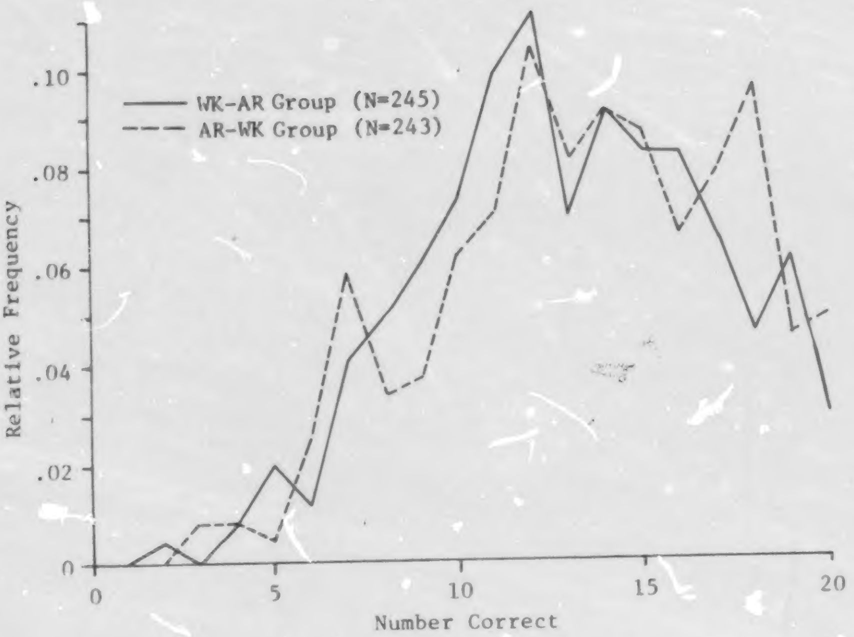
Summary Statistics for ASVAB/N, ASVAB/B, ASVAB/M, BAYES, and STMI Score Distributions with AR and WK Items, for WK-AR and AR-WK Groups

TSSM and Statistic	AR (Subgroup 1) WK-AR (N=245)	AR (Subgroup 1) AR-WK (N=243)	WK (Subgroup 2) AR-WK (N=245)	WK (Subgroup 2) WK-AR (N=247)
ASVAB/N				
Mean	12.922	13.416	22.478	22.397
S. D.	3.857	3.997	4.780	4.598
Skew	-.151	-.315	-.578	-.531
Kurtosis	-.569	-.606	-.059	-.182
Minimum	2.000	3.000	5.000	8.000
Maximum	20.000	20.000	30.000	30.000
ASVAB/B				
Mean	-.776	-.609	-.199	-.245
S. D.	1.527	1.557	1.228	1.187
Skew	-.264	-.217	-.738	-.610
Kurtosis	-.046	-.302	1.088	.950
Minimum	-5.095	-4.712	-5.248	-4.799
Maximum	2.387	2.387	2.200	2.200
ASVAB/M*				
Mean	-.536	-.328	-.067	-.078
S. D.	1.830	1.338	1.314	1.388
Skew	-.104	.247	.051	.794
Kurtosis	1.987	1.782	2.994	4.218
Minimum	-5.000	-5.000	-5.000	-5.000
Maximum	5.000	5.000	5.000	5.000
BAYES				
Mean	-.585	-.463	-.182	-.190
S. D.	1.048	1.135	.916	.869
Skew	-.248	-.545	-.021	-.161
Kurtosis	-.276	.038	-.072	.307
Minimum	-3.154	-3.899	-2.542	-2.621
Maximum	2.398	2.077	2.540	2.251
STMI*				
Mean	-.529	-.357	-.154	-.178
S. D.	1.158	1.093	.919	.871
Skew	-.892	-.632	-.003	-.063
Kurtosis	1.466	.525	-.128	.036
Minimum	-5.000	-3.814	-2.496	-2.393
Maximum	2.420	1.996	2.548	2.126

*Original boundaries were used for maximum likelihood estimates.

Figure 10
Frequency Distributions of ASVAB AR and WK Number-Correct Scores for WK-AR and AR-WK Groups

(a) AR



(b) WK

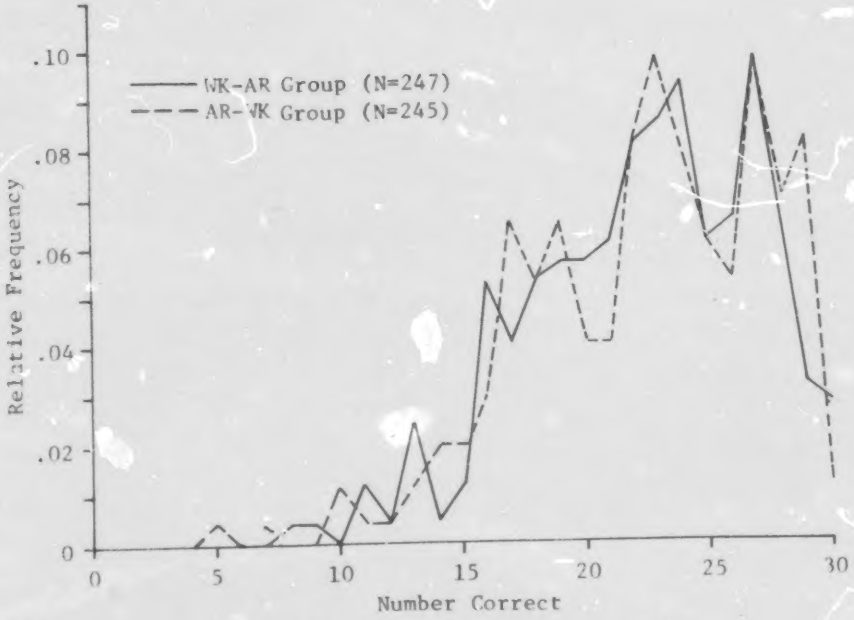
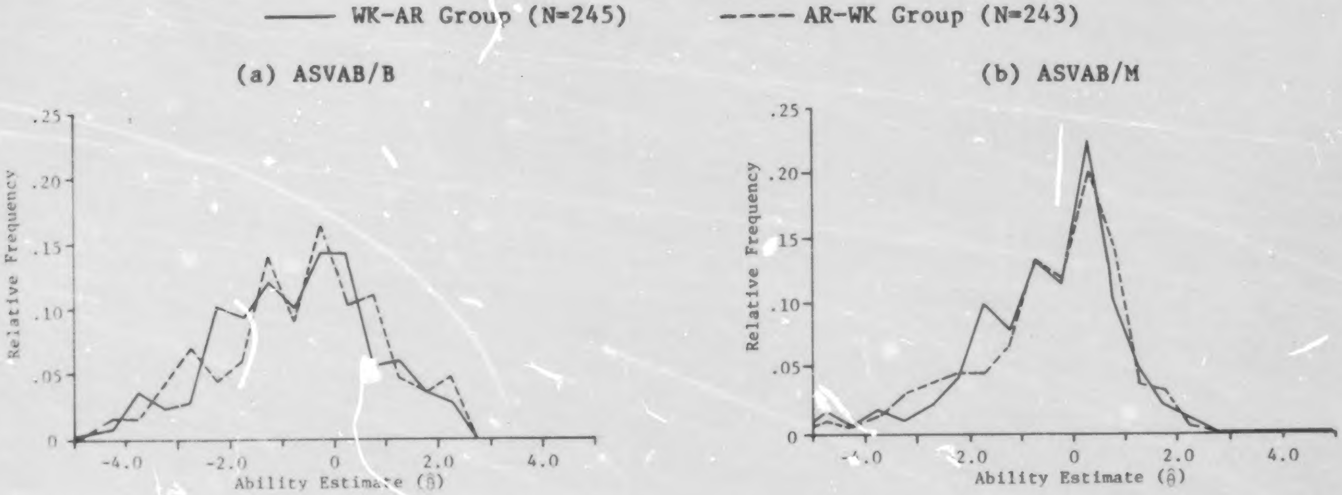


Figure 11

Frequency Distributions of AR Ability Estimates for WK-AR and AR-WK Groups



(a) ASVAB/B

(b) ASVAB/M

(c) BAYES

(d) STMI

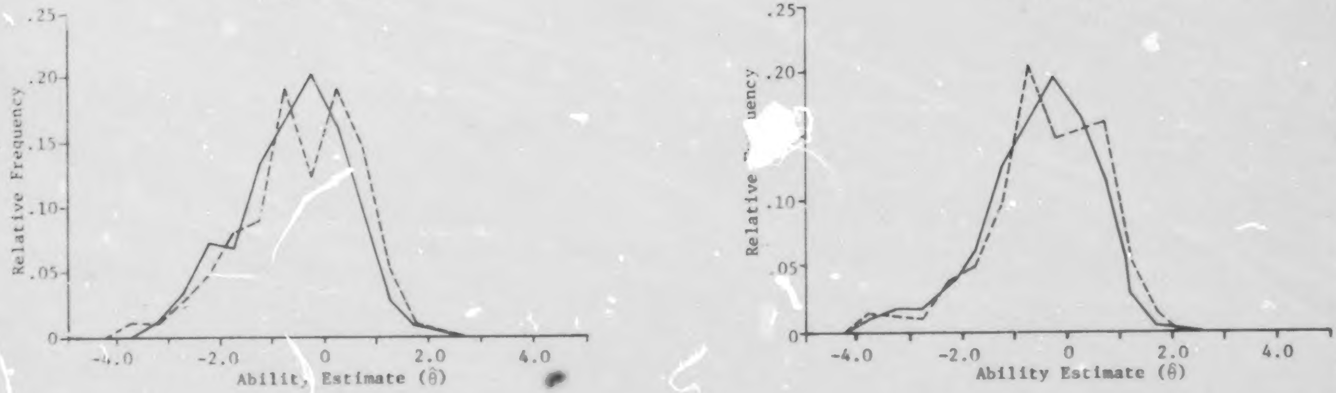
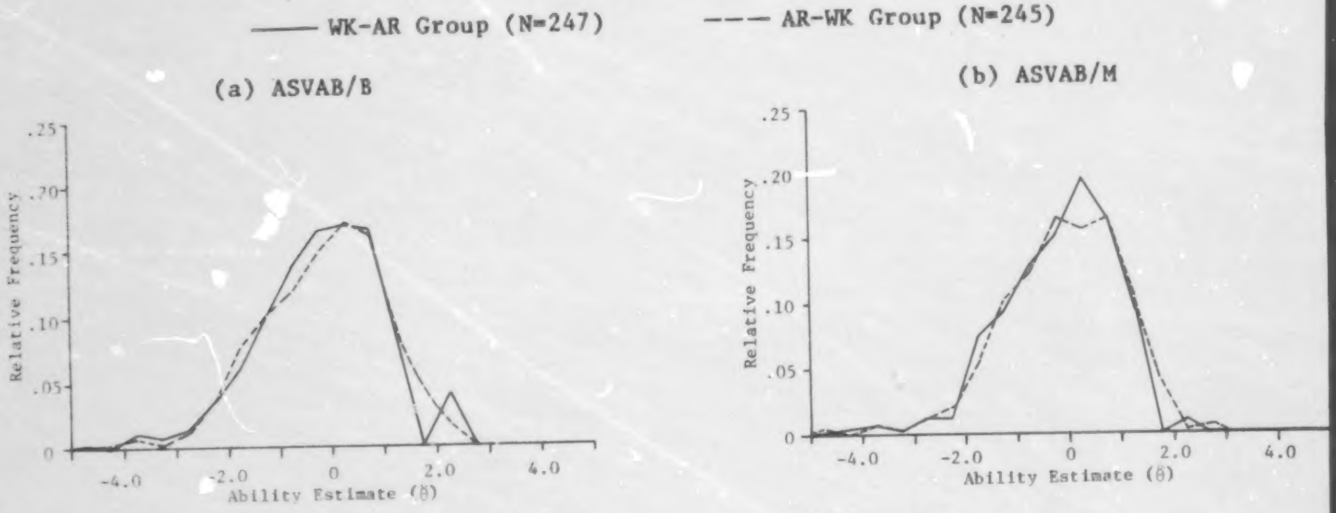
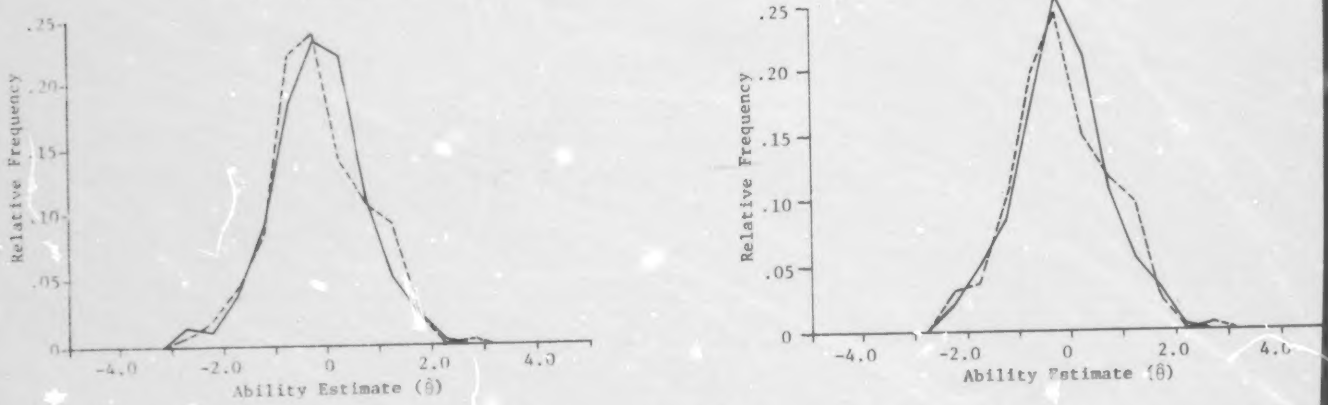


Figure 12
Frequency Distributions of WK Ability Estimates for WK-AR and AR-WK Groups



(a) ASVAB/B

(b) ASVAB/M



(c) BAYES

(d) STMI

the average ability estimate for the experimental group was about one-half standard deviation below that of the group in which the AR items were normed. There were, however, differences in the distributions of the AR ability estimates for the four TSSMs (Figure 11). The most peaked of the distributions resulted from ASVAB/M. ASVAB/B provided the least peaked distributions. The BAYES and STMI distributions were more nearly normal than the ASVAB distributions. The upper limit of the range of ability estimates was similar for the four TSSMs. However, the ASVAB IRT ability estimates contained a larger proportion of very low values (below -4.0).

Similar results were observed for the WK items (Table 13 and Figure 12). Mean ability estimates for ASVAB/B were about -.20, while for ASVAB/M the mean ability estimate was -.07. The two adaptive tests resulted in mean ability estimates closer to that of ASVAB/B than ASVAB/M. STMI resulted in the most peaked ability distribution, and the two ASVAB distributions were less peaked than the two adaptive test distributions. Again, IRT scoring of ASVAB produced more very low estimates. Since the true ability of each individual in the examinee group was, of course, unknown, it was not possible to determine which of the TSSMs resulted in estimated ability distributions which most closely approximated the underlying true ability distributions.

Correlations

Table 14 shows Pearson product-moment correlations among scores from the five TSSMs, separately for the AR and WK tests, in Subgroups 1 and 2. As expected, the data for both the AR and WK items show high correlations between ability estimates derived from BAYES and STMI, due to the fill-in procedure used in administration of the two adaptive tests. For both AR and WK items, correlations between scores on the adaptive tests and the ASVAB were lower than were scores computed from different ways of scoring the ASVAB, with a larger difference for the AR items than for the WK items.

Information

Figure 13a shows the estimated score information functions obtained for the STMI and BAYES AR adaptive tests and for ASVAB/N AR adjusted to the same length (25 items) as the adaptive tests; Figure 13b shows estimated score information functions for the two WK adaptive tests and for the ASVAB/N WK test adjusted to the same length (35 items) as the adaptive tests. The figures also show the maximum amount of information available from each adaptive-test item pool when the *k* most informative items at a θ level are administered. The amount of information available from the *k* most informative items at a θ level is an upper bound for the score information function at that θ level and provides a benchmark for evaluating the *k*-item experimental adaptive tests.

As Figure 13 shows, both of the fixed-entry adaptive tests measured all levels of ability with considerably more precision (information)

Table 14
Pearson Product-Moment Correlations Among Test Scores, Separately for AR and WK Tests, in Two Subgroups

Item Type and TSSM	TSSM 1	2	3	4	5
AR (Subgroup 1)					
1. BAYES		.922	.748	.713	.747
2. STMI	.958		.701	.670	.705
3. ASVAB/B	.796	.792		.960	.969
4. ASVAB/M	.749	.749	.958		.911
5. ASVAB/N	.785	.778	.976	.916	
WK (Subgroup 2)					
1. BAYES		.974	.833	.794	.814
2. STMI	.992		.826	.792	.810
3. ASVAB/B	.841	.840		.948	.968
4. ASVAB/M	.816	.813	.970		.895
5. ASVAB/N	.840	.836	.976	.935	

Note. Original boundaries were used for maximum likelihood estimates. For Subgroup 1, WK-AR group (N = 245) above diagonal and AR-WK group (N = 243) below diagonal. For Subgroup 2, WK-AR group (N = 247) above diagonal and AR-WK group (N =245) below diagonal.

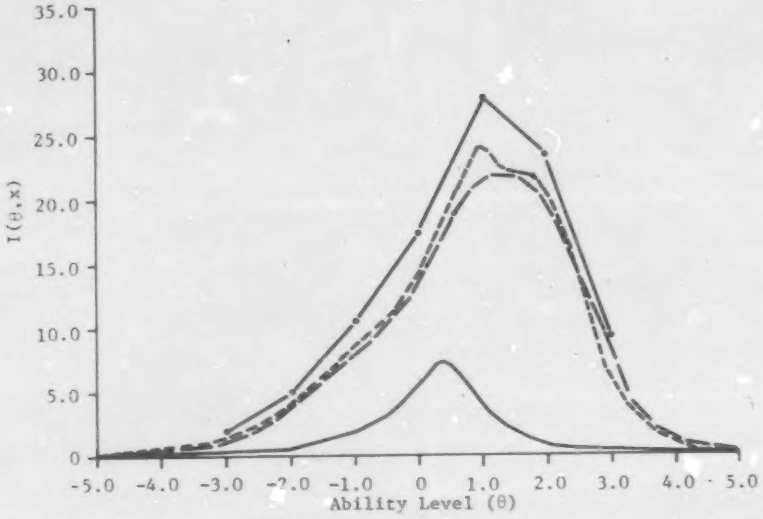
than did the ASVAB tests. For the AR tests (Figure 13a), the data show that at the maximum point of information for ASVAB/N (θ = .3) the adaptive tests provided about 2.5 times as much information as a 25-item ASVAB test. This indicates that an ASVAB AR test would need to be about 63 items long (about 3.15 times as long as the actual ASVAB-7 AR subtest) in order to measure with the same precision as the experimental adaptive tests at θ = .3. Figure 13b shows similar results for the WK tests. At the maximum level of ASVAB information (θ = .6) an ASVAB test would require about 101 items (3.37 times the length of the actual ASVAB-7 WK subtest) in order to measure as well as the 35-item experimental WK adaptive tests. Thus, the two ASVAB subtests would have to be lengthened by a factor of about 3.25 in order to approximate the level of precision available from the adaptive tests at the point of maximum information for the ASVAB tests.

Figure 13a indicates that for θ levels below about +2.0, the STMI test did slightly better at extracting information from the AR pool than did BAYES. Similarly, Figure 13b indicates that STMI was able to extract somewhat more information from the WK item pool than BAYES at most θ levels below about θ = +3.0. These results are not entirely unexpected since the STMI strategy explicitly attempts to maximize test information while the BAYES strategy does not.

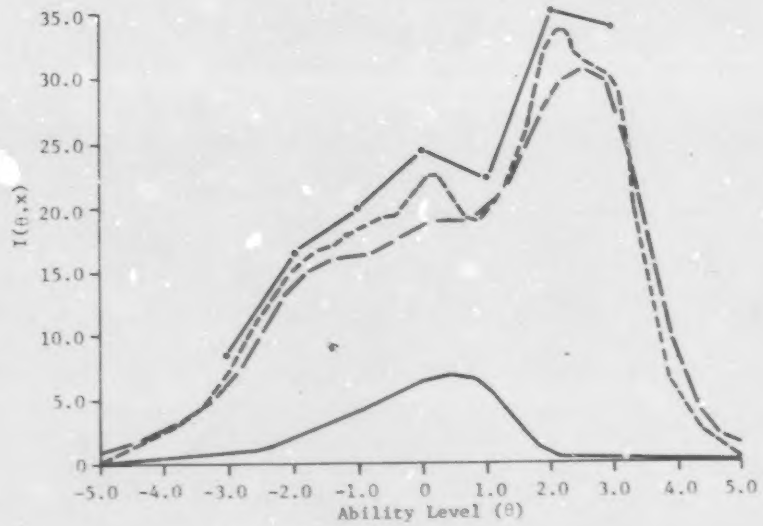
Figure 13
Score Information Functions for Fixed-Entry STMI and BAYES Tests, an ASVAB-7 Test Adjusted to the Same Length as the Adaptive Tests, and Maximum Available Information

------ STMI
—— —— BAYES
———— ASVAB/N
•—•—• Best k Items

(a) AR (k = 25 Items)



(b) WK (k = 35 Items)



It should be noted that the score information curve for any test that uses maximum likelihood estimates of θ will be influenced in the extremes of θ by the boundary values established for the maximum likelihood estimator. The boundaries of ±5.0 that were used in simulating the STMI tests no doubt tended to depress the information functions for this test at θ levels below about -3.0 and above about +3.0.

Figure 14 shows estimated score information functions for three ways of scoring the 20-item ASVAB AR test and the 30-item ASVAB WK test. As the data in Figure 14 show, little additional information was gained by scoring ASVAB by IRT methods. There were slightly larger gains for the WK subtest (Figure 14b) than for the AR subtest (Figure 14a), particularly for θ levels below about .20. For the AR subtest, differences between scoring methods occurred in the θ interval between 0 and 1.0. Differences occurred for WK for all theta values less than about 1.0, with very small differences above that value. Differences between Bayesian and maximum likelihood scoring for the two ASVAB tests were not substantial.

Validity

Single Tests

Validity as a function of test length. Sequential means, standard deviations, and validities for the adaptive tests under both FE and VE conditions are given in Appendix Tables A-1 to A-4. Figure 15 shows validity correlations obtained in the FE condition as a function of test length for the STMI and BAYES adaptive tests, and for ASVAB/N AR and WK at their normal test lengths. As the data for the AR tests show (Figure 15a) validity of ASVAB/N was about .49 at 20 items. The validity for STMI was .50 at that test length, but the validity for BAYES was about .47. For WK items (Figure 15b) the validity of ASVAB/N at its 30-item length was .29, which was equaled by BAYES, while STMI validity was about .28.

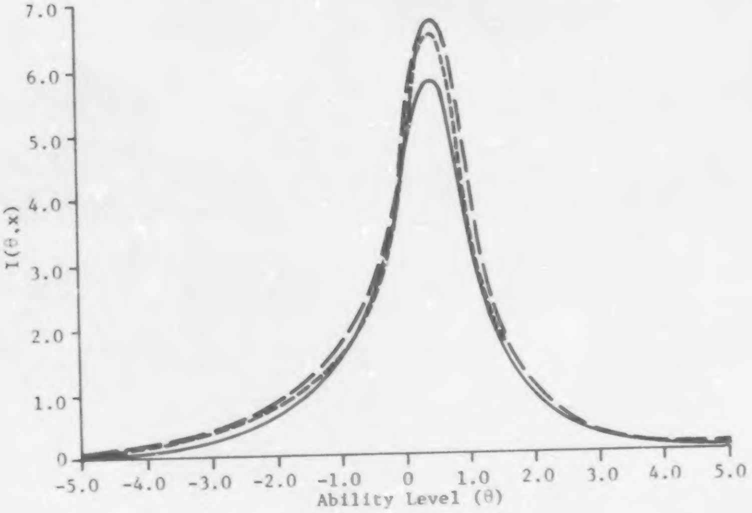
Figure 15 and Appendix Tables A-1 to A-4 show that the STMI adaptive tests reached near peak validities with relatively few items. For both item types and entry conditions, the STMI tests began to approximate their final validities at about 8 to 10 items. By contrast, the BAYES tests generally required about 20 to 22 items before validities approached their final values. Under fixed-entry conditions, the STMI AR and WK validities at 8 to 10 items were near those of the 20- and 30-item ASVAB/N tests. Under variable-entry conditions, STMI validities at 8 to 10 items were still somewhat below the ASVAB/N validities.

Effect of maximum likelihood boundary values. The sequential validity data shown in Figure 15 for STMI were computed with the original bounds for maximum likelihood ability estimates. Subsequent to those analyses, virtually all remaining validity analyses were conducted using a set of revised bounds for maximum likelihood estimates. Table 15

Figure 14
Score Information Functions for Three Scores on ASVAB-7 Subtests

— — Bayesian
---- Maximum Likelihood
—— Number Correct

(a) AR



(b) WK

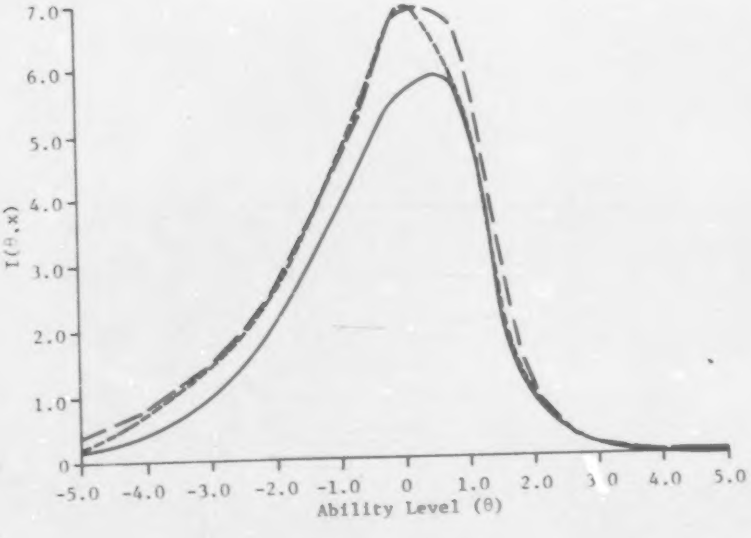
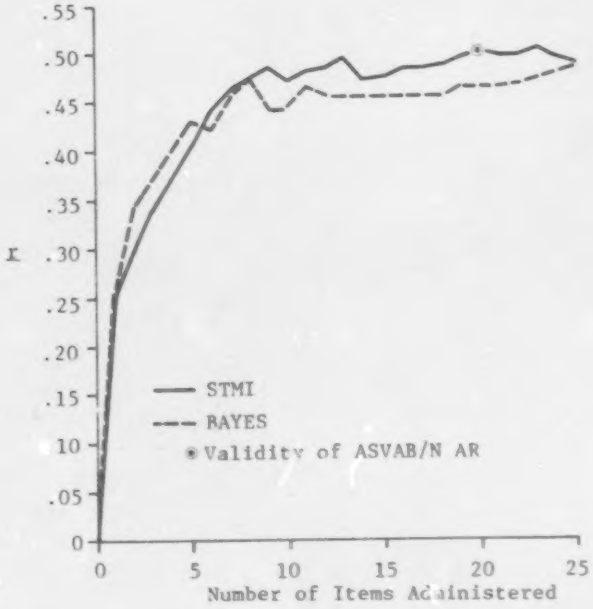
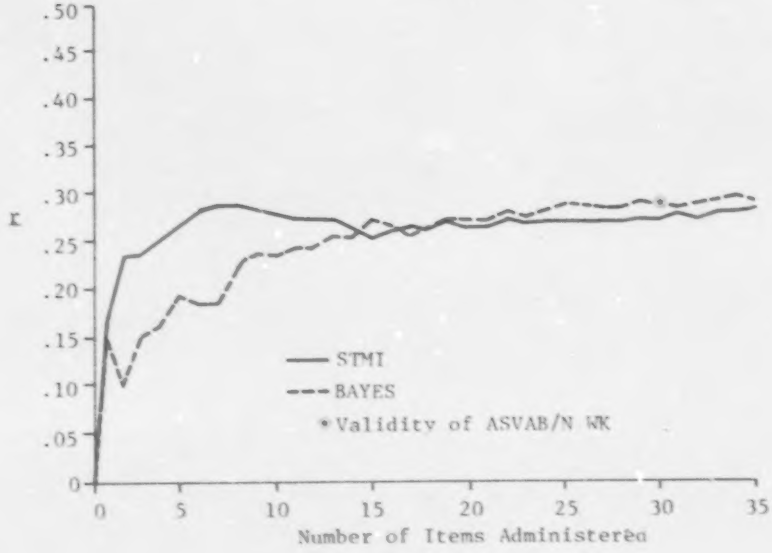


Figure 15
Criterion-Related Validity Correlations of Fixed-Entry Adaptive Tests as a Function of Number of Items Administered, and Validity of ASVAB/N

(a) AR (AR-WK Group, N=221)



(b) WK (WK-AR Group, N=231)



shows the effect on terminal validity (i.e., after 20 AR items and 30 WK items) of using the original bounds, in-bounds cases only, and the revised bounds described previously. As these data show, for STMI the use of in-bounds estimates resulted in one increase and five decreases in criterion-related validity, whereas the use of revised bounds resulted in no change in four instances and slight increases in validity in two instances. For ASVAB/M, in-bounds-only estimates resulted in four increases in validity, but in decreases for AR items. On the other hand, the revised bounds for ASVAB/M resulted in increases in validity in all cases. Consequently, all of the following analyses (with one exception) were conducted with the revised bounds.

Table 15
Criterion-Related Validity Correlations of Maximum Likelihood Ability Estimates for Original-Bounds, In-Bounds, and Revised-Bounds Scoring

Test and Group	Original Bounds[a]	In-Bounds Only[b]	Revised Bounds[a]
STMI AR			
AR-WK	.501	.466	.501
WK-AR	.406	.346	.411
STMI WK			
AR-WK	.346	.302	.346
WK-AR	.276	.294	.276
STMI Composite			
AR-WK	.534	.502	.534
WK-AR	.459	.426	.461
ASVAB/M AR			
AR-WK	.425	.402	.443
WK-AR	.396	.357	.407
ASVAB/M WK			
AR-WK	.370	.373	.389
WK-AR	.250	.337	.279
ASVAB/M Composite			
AR-WK	.490	.497	.516
WK-AR	.421	.438	.442

[a] N = 221 in AR-WK group and N = 231 in WK-AR group.
[b] N = 201 in AR-WK group and N = 208 in WK-AR group.

Linear-model analysis. Table 16 shows the criterion-related validity correlations for each TSSM under each order of administration, separately for the AR and WK items. The grand mean of the validity correlations across all of the 20 conditions was .392, with a standard deviation of .072. Table 17 summarizes the marginal means collapsed across the order conditions separately for the AR and WK items, and the combined marginal mean for each testing strategy across order conditions and item types.

Table 16
Criterion-Related
Validity Correlations
for Single Tests

TSSM and Order	Item Type	Cell Code	r
BAYES			
1	AR	1	.467
2	AR	2	.457
1	WK	3	.294
2	WK	4	.347
STMI*			
1	AR	5	.501
2	AR	6	.411
1	WK	7	.276
2	WK	8	.346
ASVAB/B			
1	AR	9	.456
2	AR	10	.443
1	WK	11	.292
2	WK	12	.399
ASVAB/M*			
1	AR	13	.443
2	AR	14	.407
1	WK	15	.279
2	WK	16	.389
ASVAB/N			
1	AR	17	.491
2	AR	18	.448
1	WK	19	.294
2	WK	20	.409
Mean			.392
S.D.			.072

*Revised boundaries were used for maximum likelihood estimates.

For the AR items, the data show highest mean validity for ASVAB/N (.469; see Table 17) with BAYES resulting in a mean validity of .462. Lowest mean validity was for ASVAB/M (.425), with ASVAB/B intermediate between the lowest and the highest ASVAB mean validities (.449). STMI for the AR items had a mean validity of .456. For the AR items there appeared to be a slight order effect with higher validities for all TSSMs occurring in the Order 1 condition (Table 16).

For WK items Table 17 shows the highest mean validity again for ASVAB/N (.352). ASVAB/B obtained second highest mean validity (.345),

Table 17
Marginal TSSM Means for Single-Test Validities

TSSM	Marginal Mean AR	Marginal Mean WK	Marginal Mean Combined
BAYES	.462	.320	.391
STMI*	.456	.311	.383
ASVAB/B	.449	.345	.397
ASVAB/M*	.425	.334	.380
ASVAB/N	.469	.352	.410

*Revised boundaries were used for maximum likelihood estimates.

followed by ASVAB/M (.334). Both the adaptive tests obtained lowest mean validities for the WK items, with BAYES resulting in a mean validity correlation of .320, and STMI the lowest with $\underline{r} = .311$. For the WK items, the apparent order effect was reversed, with highest validities for all TSSMs in the Order 2 condition (Table 16). The largest difference occurred for the ASVAB/N condition (cell codes 19 and 20) where the Order 1 condition obtained a validity of .294, while the validity in the Order 2 condition was .409.

Table 18
Three-Way Linear-Model Analysis for Single-Test Validities

Effect	df_1	df_2	F	$\underline{p}$	Proportion of Variance
Main Effects					
TSSM (A)	4	222	3.02*	.019	.023
Item Type (B)	1	225	6.67*	.010	.688
Order (C)	1	225	.25	.619	.033
2-Way Interactions					
A X B	4	222	.83	.505	.021
A X C	4	222	1.61	.174	.019
B X C	1	225	1.19	.277	.200
3-Way Interaction	4	222	3.02*	.019	.016

*Statistically significant at $\underline{p} \leq .05$.

Table 18 shows the results of the three-way linear-model analysis for single-test validities. As the data show, there was a significant main effect for TSSM, a significant main effect for Item Type, and a significant three-way interaction among the independent variables. Figure 16 plots the validity correlations shown in Table 16, to illustrate the nature of the three-way interaction. Table 19 shows the results of significance tests for the three-way interaction contrasts, each with a single degree of freedom. The only significant contrast was the one

Figure 16
Three-Way Interaction of TSSM, Item Type, and Order
(Cell Codes Are in Parentheses)

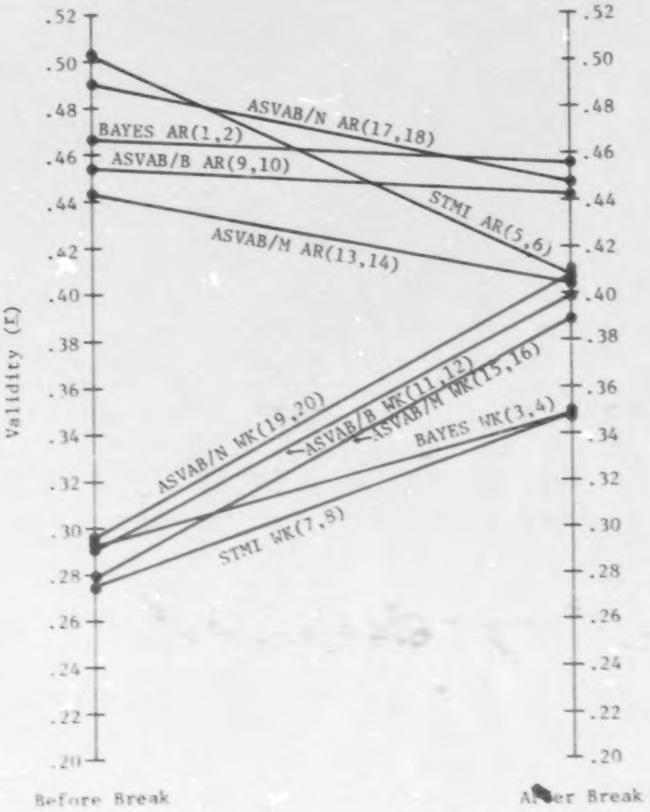


Table 19
Three-Way Interaction Contrasts for Single-Test Validities

Cells Contrasted	F	p
[(1+4) - (2+3)] - [(17+20) - (18+19)]	1.39	.239
[(5+8) - (6+7)] - [(17+20) - (18+19)]	.00	.968
[(9+12) - (10+11)] - [(17+20) - (18+19)]	1.80	.182
[(13+16) - (14+15)] - [(17+20) - (18+19)]	.13	.715
[(1+4) - (2+3)] - [(13+16) - (14+15)]	1.05	.308
[(5+8) - (6+7)] - [(13+16) - (14+15)]	.05	.823
[(9+12) - (10+11)] - [(13+16) - (14+15)]	1.36	.245
[(1+4) - (2+3)] - [(9+12) - (10+11)]	.50	.478
[(5+8) - (6+7)] - [(9+12) - (10+11)]	.30	.585
[(1+4) - (2+3)] - [(5+8) - (6+7)]	8.56*	.004

Note. 1 and 225 degrees of freedom for all contrasts.
*Statistically significant at $p \leq .005$.

involving cell codes 1+4 and 2+3 versus 5+8 and 6+7. As shown in Figure 16, this contrast involved the BAYES tests and the STMI tests. For these two adaptive tests, there was a differential Item Type by Order two-way interaction effect. For STMI, the difference between validities obtained in the AR-WK condition and those obtained in the WK-AR condition was .08. For BAYES, this same difference was approximately .03. Thus, STMI was more sensitive to the Item Type by Order two-way interaction. None of the three-way interaction contrasts involving the conventional tests were statistically significant.

As shown in Table 18, the marginal Item Type by Order two-way interaction was not significant in spite of the fact that this effect accounted for 20 percent of the variance in single-test validities. This is because the contrast involved in testing this effect is a between-groups contrast. As mentioned earlier, this results in a test that is less sensitive than the other tests in the linear-model analysis. Nevertheless, the presence of this two-way interaction is clearly seen in Figure 16 where the slopes of all lines connecting AR tests are negative and the slopes of all lines connecting WK tests are positive. As shown in Table 16 and Figure 16, the significant main effect for Item Type was due to the tendency for AR tests to have higher validities than WK tests, regardless of TSSM or order condition.

Table 20
Pairwise Contrasts Among
Marginal TSSM Means
for Single-Test Validities

Contrast	F	p
BAYES vs. ASVAB/N	.90	.343
STMI vs. ASVAB/N	1.55	.215
ASVAB/B vs. ASVAB/N	3.22	.074
ASVAB/M vs. ASVAB/N	9.55*	.002
BAYES vs. ASVAB/M	.37	.545
STMI vs. ASVAB/M	.07	.798
ASVAB/B vs. ASVAB/M	10.20*	.002
BAYES vs. ASVAB/B	.09	.770
STMI vs. ASVAB/B	.38	.536
BAYES vs. STMI	.64	.424

Note. 1 and 225 degrees of freedom for all contrasts.
*Statistically significant at $p \leq .005$.

Table 20 shows the results of testing pairwise contrasts among the marginal TSSM means. There were two significant contrasts, both involving only ASVAB tests. The mean ASVAB/N validity of .410 (see Table 17) was significantly higher than the mean ASVAB/M validity of .380. ASVAB/B also had a significantly higher mean validity correlation (.397) than ASVAB/M. None of the differences in validities between the adaptive and conventional tests were statistically significant.

Composites

Intercorrelations between AR and WK scores. Table 21 shows the intercorrelations among the 10 test scores generated for members of Subgroup 4. In addition to being computed in a different subgroup than the correlations in Table 14, these data are based on the revised boundaries for maximum likelihood ability estimates. The AR and WK cross-correlations in Table 21 suggest a tendency for WK and AR scores to correlate lower for the two adaptive tests than they did for the ASVAB tests. In particular, the cross-correlations between AR and WK scores for BAYES and STMI were .172 and .158, respectively, in the WK-AR group and .288 and .285 in the AR-WK group. By contrast, ASVAB/N AR and WK scores correlated .313 in the first group and .333 in the second group.

Table 21
Pearson Product-Moment Correlations Among Test Scores in the WK-AR and AR-WK Graduate Groups, for AR and WK Items

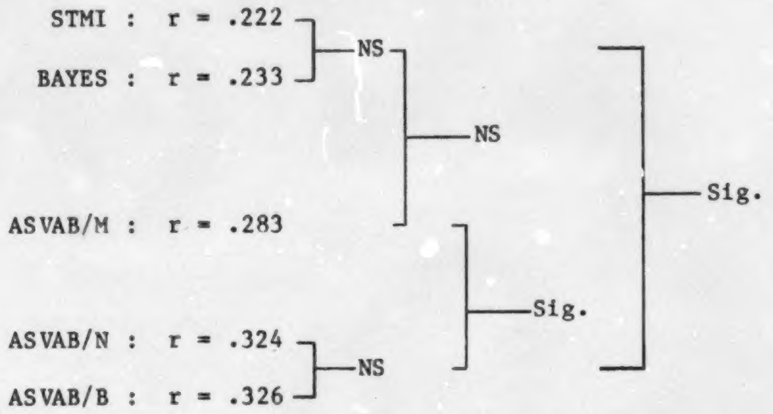
Item Type and TSSM	AR					WK				
	1	2	3	4	5	1	2	3	4	5
AR										
1. BAYES		.930	.750	.717	.744	.172	.191	.192	.187	.180
2. STMI*	.953		.714	.687	.712	.140	.158	.150	.143	.134
3. ASVAB/B	.784	.783		.973	.968	.298	.301	.311	.298	.318
4. ASVAB/M*	.768	.772	.978		.928	.260	.261	.268	.255	.272
5. ASVAB/N	.771	.768	.974	.948		.287	.287	.295	.278	.313
WK										
1. BAYES	.288	.301	.300	.275	.277		.973	.826	.821	.801
2. STMI*	.270	.285	.285	.260	.263	.993		.819	.817	.799
3. ASVAB/B	.302	.315	.338	.311	.329	.833	.834		.990	.966
4. ASVAB/M*	.300	.315	.338	.309	.328	.834	.835	.991		.948
5. ASVAB/N	.313	.326	.343	.322	.333	.832	.830	.975	.961	

Note. Correlations in WK-AR Group (N = 231) appear above the main diagonal. Correlations in AR-WK Group (N = 221) appear below the main diagonal.

*Revised boundaries were used for maximum likelihood estimates.

Figure 17 shows the results of significance tests of the differences among the cross-correlations obtained after pooling the AR-WK and WK-AR groups in Subgroup 4. As Figure 17 shows, there were no significant differences in the cross-correlations for STMI, BAYES, and ASVAB/M. ASVAB/M, with r = .283 was significantly different ($p \leq .05$) than ASVAB/N and ASVAB/B (r = .324 and .326). Finally, there were significant differences between the cross-correlations obtained for both STMI and BAYES and the cross-correlations obtained for ASVAB/N and ASVAB/B. In all cases, the cross-correlations were significantly lower ($p \leq .05$) for the adaptive tests than they were for the ASVAB tests.

Figure 17
Cross-Correlations Between Test Scores
from AR and WK Tests (N=452)



Linear-model analyses. Table 22 shows criterion-related validity correlations for both the fixed-weight and optimally-weighted composites, separately for each TSSM and content-order group. For both fixed- and optimally-weighted composites, there appeared to be a difference in validity correlations between the AR-WK group and the WK-AR group. In each case, the first group obtained higher validity correlations than did the second group. For the fixed-weight composites, the highest single correlation was .545 for ASVAB/N, while the highest average correlation with content-order groups combined was .506 for BAYES. For the optimally-weighted composites the highest single validity correlation was .555 for ASVAB/N, and the highest average correlation across content-order groups combined was .515 for ASVAB/N, with BAYES obtaining a mean correlation of .511. Lowest average validity correlations were .479 for ASVAB/M with fixed weights, and .481 for ASVAB/M with optimal weights. Average validities for STMI and ASVAB/B were all near .50.

It is interesting to note that while there was a mean validity difference of .023 in favor of ASVAB/N over BAYES and STMI for single-test validities (Table 17), there was a mean validity difference of .005 in favor of BAYES and STMI using fixed-weight composites and a mean validity difference of only .008 in favor of ASVAB/N for the optimally-weighted composites. Thus, mean adaptive-test validities were essentially equal to those of ASVAB/N when composite scores were computed from the single-test scores. This was a result of the lower cross-correlations between AR and WK scores for the two adaptive tests (average cross-correlation = .226 in Table 21) in comparison to ASVAB/N (average cross-correlation = .323 in Table 21).

Table 23 shows the results of the two-way linear-model analyses for both fixed-weight and optimally-weighted composites. For the fixed-weight composites there was a significant two-way interaction, but no

Table 22
Criterion-Related Validity Correlations for Fixed-Weight and Optimally-Weighted Composites

TSSM and Content Order	Cell Code	Fixed Weights	Optimal Weights
Bayes			
AR-WK	1	.513	.517
WK-AR	2	.500	.506
Mean		.506	.511
STMI*			
AR-WK	3	.534	.543
WK-AR	4	.461	.463
Mean		.498	.503
ASVAB/B			
AR-WK	5	.525	.525
WK-AR	6	.464	.472
Mean		.494	.498
ASVAB/M*			
AR-WK	7	.516	.517
WK-AR	8	.442	.445
Mean		.479	.481
ASVAB/N			
AR-WK	9	.545	.555
WK-AR	10	.451	.476
Mean		.498	.515

*Revised boundaries were used for maximum likelihood estimates.

Table 23
Two-Way Linear-Model Analyses for Fixed-Weight-Composite and Optimally-Weighted-Composite Validities

Type of Weight and Effect	df_1	df_2	F	p	Proportion of Variance
Fixed Weights					
Main Effects					
TSSM	4	222	1.21	.307	.060
Content Order	1	225	.90	.345	.790
2-Way Interaction	4	222	3.05*	.018	.150
Optimal Weights					
Main Effects					
TSSM	4	219	1.35	.254	.125
Content Order	1	222	.35	.556	.731
2-Way Interaction	4	219	1.81	.128	.144

*Statistically significant at $p \leq .05$.

significant main effects. For the optimally-weighted composites there were no significant differences among the validities.

Table 24 shows the results of a posteriori significance tests for the significant two-way interaction between TSSM and Content Order for fixed-weight composites. As the data show, the contrast between cell codes 1 and 2 versus 3 and 4 approached statistical significance. This suggests that the significant two-way interaction resulted from the lack of a Content Order effect for BAYES in contrast to the substantial Content Order effect for STMI (see Table 22). None of the contrasts involving the adaptive versus the conventional tests were statistically significant. The absence of a significant two-way interaction for optimally-weighted composites presumably reflects the fact that estimating regression weights instead of using fixed weights tends to increase the sampling variance of the contrasts tested.

Table 24
Two-Way Interaction Contrasts for
Fixed-Weight-Composite Validities

Cells Contrasted	F	p
(1-2) - (9-10)	2.76	.098
(3-4) - (9-10)	.17	.680
(5-6) - (9-10)	2.71	.101
(7-8) - (9-10)	.68	.411
(1-2) - (7-8)	1.50	.223
(3-4) - (7-8)	.00	.978
(5-6) - (7-8)	.73	.393
(1-2) - (5-6)	1.03	.311
(3-4) - (5-6)	.07	.786
(1-2) - (3-4)	7.74*	.006

Note. 1 and 225 degrees of freedom for all contrasts.
*Approaches statistical significance at $p \leq .005$.

Comparison with pre-enlistment ASVAB composites. Table 25 shows criterion-related validity correlations computed in the two content-order groups and the combined content-order group in Subgroup 5. Original boundaries were used for the maximum likelihood ability estimates. The table also shows validity correlations for five pre-enlistment ASVAB composites in this subgroup. The most informative comparisons in the table are between the pre-enlistment General-Technical composite (WK+AR) and the five experimental WK+AR composites.

Since the pre-enlistment ASVAB tests were composed of items presented in a standard booklet format, comparison of the average experimental-composite validity in the combined group (.491 over all TSSMs) with the pre-enlistment General-Technical-composite validity of .493 in

Table 25
Criterion-Related Validity of Pre-Enlistment
ASVAB Composite Scores and Experimental
Fixed-Weight-Composite Scores

Composite	Group WK-AR (N=206)	 AR-WK (N=200)	 Combined (N=406)
Pre-Enlistment ASVAB			
Mechanical	.381	.368	.377
Administative	.227	.332	.281
General-Technical	.478	.510	.493
Electronics	.414	.429	.421
AFQT	.507	.495	.500
Experimental			
BAYES	.499	.494	.497
STMI*	.451	.524	.487
ASVAB/B	.478	.532	.504
ASVAB/M*	.433	.494	.462
ASVAB/N	.463	.551	.506

*Original boundaries were used for maximum likelihood estimates.

this same group suggests that computer administration of aptitude test items does not degrade the validity of scores derived from these items. The only experimental composite that differed substantially from the pre-enlistment General-Technical composite in the combined group was ASVAB/M with a validity of .462. This difference was presumably an effect due to scoring method since the other two experimental-ASVAB-composite validities actually exceeded the pre-enlistment General-Technical composite's validity in the combined group.

The difference of .032 between the validity of the pre-enlistment General-Technical composite in the AR-WK group and the validity of this composite in the WK-AR group suggests that the AR-WK group may have been a "more predictable" sample even when AR and WK items were given to these two groups in the same order (WK before AR in the pre-enlistment ASVAB). This suggests that a portion of the non-significant two-way-interaction effect that was seen in Figure 16, and also the group differences mentioned in the discussion of Table 22, might be attributable to simple randomization error.

DISCUSSION AND CONCLUSIONS

This study provided t[illegible]st direct comparison between the STMI and Bayesian adaptive testing strategies using the same examinees and the same item pools for both tests. Results indicated that the two testing strategies tended to select the same items for most individuals.

On the average, about 85% of the items selected for an individual by one adaptive strategy were also selected by the other strategy.

Two factors contributed to this finding. First, both the item information function, which is used by STMI in selecting items, and the expectation of the Bayesian posterior variance, which is minimized by BAYES during item selection, are strongly influenced by an item's discrimination parameter (a). Thus, items having high a values will tend to be selected by both strategies. Second, the BAYES strategy used relatively diffuse prior distributions with standard deviations of 2.15 and 2.35 (in the fixed-entry condition), or 1.76 and 1.96 (in the variable-entry condition) for AR and WK, respectively. This allowed relatively large fluctuations in the Bayesian ability estimates during testing and this, in turn, resulted in the selection of items that would not have been used by the BAYES strategy if the prior standard deviation had been set to 1.0, as it frequently is.

The data on computer response time showed that the Bayesian adaptive test took considerably more computer time than did the STMI or ASVAB tests, which were very similar in computer response time. The observed difference between the STMI and BAYES adaptive tests reflects the fact that STMI did not require a search of the full item pool in order to select each test item. Many of the computer response times observed for the Bayesian adaptive tests were longer than desirable in an operational testing environment. However, these computer response times were largely dependent on the characteristics of the computer system utilized and the design of the operating system software. Subsequent experience with the design of similar adaptive testing systems, using more sophisticated computer hardware and improved operating system software, has shown that the long computer response times observed in this study for the Bayesian strategy can be reduced to acceptable response times of two seconds or less. Hence, these data should not be viewed as detrimental to the application of the Bayesian strategy in operational testing environments.

The results of this study were consistent with earlier studies in showing longer examinee response times for adaptive versus conventional tests (Johnson, Weiss, and Prestwood, 1981; Martin et al., in press; Waters, 1977). In this study, as in the earlier studies, the differences in examinee response times were largely a function of the differing item difficulties of the adaptive and conventional tests. Since the adaptive tests tailored item difficulty to each individual's ability level, most examinees received more difficult items in the adaptive testing condition than in the conventional testing condition. The ASVAB subtests contained many items that were too easy for the examinee sample. Consequently, the time required by an examinee to answer the ASVAB items was usually less than it was for the items in the adaptive tests. The data did suggest that for a given amount of testing time, even if fewer items are administered in an adaptive test, levels of measurement precision that significantly exceed the precision available from standard ASVAB tests are readily attainable.

The results of the comparison between the fixed-entry and variable-entry testing conditions were equivocal. For three of the four adaptive tests the possibility of a slight advantage in favor of variable entry was indicated. However, the observed differences were not large and for the STMI AR test there was a substantial difference in favor of the fixed-entry condition. The STMI AR correlations between interim and final ability estimates under variable entry were found to be lower at each point in the test than similar correlations computed for the other STMI and BAYES tests regardless of item type (AR or WK) and entry type. The reason for this finding was not immediately apparent.

Score Information

The results of this study agreed with previous research in demonstrating higher levels of information/precision for the adaptive tests in comparison to the conventional tests. The information analysis showed that the ASVAB subtests would have to be increased in length substantially to equal the precision of the adaptive tests.

The information functions shown in Figure 14 indicate that the substantial differences seen in Figure 13 between ASVAB and the adaptive tests are not a result of the fact that ASVAB/N, rather than ASVAB/B or ASVAB/M, was used for the comparison. In fact, the real reason for the large differences observed in Figure 13 is to be found in Tables 2 and 3. In Table 3 it is seen that both ASVAB subtests contained a number of items that were really too easy to provide much information near $\theta = 0$. Moreover, the median level of item discrimination in the ASVAB subtests was rather low (about .70) and the median c parameter was near .25. On the other hand, Table 2 shows that both adaptive-test item pools contained a large number of items with difficulties that provided adequate information near $\theta = 0$ and that the median a and c parameters in these pools were about .90 and .18, respectively.

Since adaptive tests systematically select items with higher than average levels of discrimination and lower than average c parameters, it may be assumed that the adaptive testing strategies improved on the already obvious superiority of their item pools. A rough comparison can be made by assuming that each adaptive test selected items falling in the upper 30% of each item pool's a distribution and the lower 30% of each pool's c distribution. Given this assumption, the data in Table 2 suggest that the median a among items actually administered in the adaptive tests may have been about 1.25 (roughly the 80th percentile of the two a distributions).

Since item information, which has an indirect effect on score information, is an increasing function of a^2 (Lord, 1980, p. 73) the reason for the substantial differences observed in Figure 13 becomes obvious. Squaring the median ASVAB a (.70) and the roughly approximated median adaptive-test a (1.25) gives .49 and 1.56, respectively. The ratio of the latter to the former (3.19) approximates the proportional increase in length required in order for the two ASVAB subtests to obtain information levels comparable to the adaptive tests at $\theta = .3$ (for AR) and $\theta = .6$ (for WK).

The ASVAB items were found to be poorly suited for the task of precise measurement near the mean of the Air Force enlistee population. While conventional tests could be constructed that would provide higher levels of information than the ASVAB subtests by selecting highly discriminating items from the adaptive-test item pools, and such tests might even provide somewhat more information than the adaptive tests over short, selected θ intervals, they could not provide the same high level of information over a wide range of ability.

Validity

Validity analyses at the single-test level showed a small nonsignificant advantage in favor of the conventional tests over the adaptive tests. However, when both equally-weighted and optimally-weighted composite scores were computed, the adaptive testing strategies provided virtually identical average levels of criterion-related validity. This finding was attributed to the tendency for the adaptive strategies to generate lower cross-correlations between individual AR and WK scores.

It is possible that by selecting more highly discriminating items, adaptive tests tend to generate ability estimates that are more nearly "factor pure." If this conjecture is correct, it would account for the lower adaptive-test cross-correlations observed in this study.

Pairwise contrasts among marginal TSSM mean validities for single tests indicated that maximum likelihood scoring of the ASVAB subtests resulted in significantly lower levels of validity than either number-correct scoring or Bayesian scoring of these tests. None of the contrasts between the adaptive tests and the various scorings of ASVAB were statistically significant.

A significant interaction involving STMI and BAYES was observed in both the linear-model analysis for single tests and the linear-model analysis for equally-weighted composites. In both cases, STMI was significantly more sensitive to the tendency for AR validities to be lower when the AR items were administered in the second half of the testing session. It seems likely that there is a connection between these significant interactions and the finding that interim and final ability estimates correlated less highly for the STMI AR test under the variable-entry condition (which was implemented in the second half of each testing session) than for any other combination of adaptive testing strategy, item type, and entry type. Further research is needed to identify the source of these apparently related effects.

The data on adaptive-test validity as a function of test length showed that the STMI adaptive tests approached their terminal validities after only 8 to 10 items. Moreover, under the fixed-entry condition, STMI validities at 8 to 10 items approximated the validities of the much longer ASVAB subtests. The Bayesian adaptive tests were observed to approach their terminal validities somewhat more slowly than did STMI.

Relationship with Previous Research

Four previous studies of the validity of computer-administered adaptive tests were briefly discussed above in the introduction. Table 26 summarizes some of the results obtained in three of these studies and in one other study (Johnson & Weiss, 1980) that examined only the alternate-forms reliabilities of conventional and adaptive tests.

Table 26
Alternate-Forms Reliability and Concurrent-Validity Correlations at a Test Length of 30 Items from Four Studies of Adaptive Testing

	Alternate-Forms Reliability		Concurrent Validity	
Study	Conventional Test	Bayesian Adaptive Test	Conventional Test	Bayesian Adaptive Test
Kingsbury & Weiss (1980)	.88	.92	.84	.80
Johnson & Weiss (1980)	.90	.81	**	**
McBride (1980)	.88*	.90*	.87	.85
Martin et al. (in press)	.89	.90	.81	.84

*Revised values provided by McBride, personal communication.
**Validity coefficients were not computed in this study.

All four of the studies listed in Table 26 utilized a 30-item Bayesian adaptive word-knowledge test as one of the testing strategies studied. Each study compared the Bayesian adaptive test to a 30-item conventional word-knowledge test. (Johnson and Weiss also studied a 30-item maximum-information strategy somewhat like the STMI method, but results for this test will not be presented here since none of the other studies used this strategy.) Both Kingsbury and Weiss (1980) and Johnson and Weiss (1980) tested college students and used a conventional predictor test with a peaked distribution of item difficulties. McBride (1980) and Martin et al. (in press) tested Marine recruits and used a conventional predictor test with a broad range of item difficulties. In spite of these differences between the studies, there is a remarkable degree of consistency among the four conventional-test alternate-forms reliability coefficients shown in Table 26. The values range from .88 to .90, with a mean of .89. The adaptive-test alternate-forms reliabilities obtained in three of the studies are also quite similar, ranging from .90 to .92 with a mean of .91. However, the value obtained by Johnson and Weiss (.81) deviates substantially from the others.

A likely explanation for the anomalous result obtained by Johnson and Weiss is found in their description of the method used to assemble their adaptive-test item pool. They state that "Items with discrimination parameters of $\underline{a} = 3.00$ were routinely rejected because this value was identified as a statistical artifact of the parameterization program and not as a true reflection of the item's discrimination value" (p. 19). While the limiting value imposed on item discrimination parameters during calibration of the item bank that Johnson and Weiss had available for item-pool construction was admittedly arbitrary, items that attained this limiting value should have been included in Johnson and Weiss' adaptive-test item pool. As a group, these items would have been among the best items available in the item pool. Eliminating them from consideration served to reduce the adaptive test's discriminating power and, presumably, its reliability.

Support for this possible explanation of Johnson and Weiss' results is found in the adaptive-test reliability coefficient obtained by Kingsbury and Weiss (.92 in Table 26). The adaptive-test item pool assembled by Kingsbury and Weiss consisted of items drawn from the same large item bank that was available to Johnson and Weiss. However, Kingsbury and Weiss did not exclude items with estimated $\underline{a}$ values of 3.00. In fact, examination of Kingsbury and Weiss' Appendix Table C reveals that 26% of the items in their adaptive-test item pool had estimated $\underline{a}$ values at the limiting boundary value of 3.00. The median estimated $\underline{a}$ value among the items selected by Kingsbury and Weiss for their adaptive-test item pool was 1.20. This may be contrasted with the mean estimated $\underline{a}$ value of .76 in the Johnson and Weiss adaptive-test item pool (Waters, 1980, p. 52). Since the Bayesian adaptive test implemented by Johnson and Weiss was handicapped by the limitations of its item pool, the adaptive-test alternate-forms reliability obtained by Johnson and Weiss should not be considered comparable to the other three values appearing in column two of Table 26.

Thus, on the basis of the available empirical evidence, it seems reasonable to conclude that a 30-item Bayesian adaptive word-knowledge test will tend to be somewhat more reliable (by about .02) than a typical 30-item conventional test. This conclusion is consistent with the earlier discussion of the tendency for adaptive tests to have rather high score information functions over a fairly wide range of ability.

Columns three and four of Table 26 present average "concurrent-validity" coefficients for the 30-item conventional and Bayesian-adaptive tests in three of these studies. Each of the entries in columns three and four of Table 26 is the average of two concurrent-validity coefficients. In the Kingsbury and Weiss study, the "criterion" was a 120-item Bayesian-scored conventional word-knowledge test. In the McBride and Martin et al. studies the criterion was a 50-item number-correct-scored conventional test.

The Kingsbury and Weiss study is unique in that it used a large sample ($N = 472$) and a repeated measures design. This experimental design served to eliminate the effect of between-groups sampling error from comparisons between the adaptive and conventional tests. The McBride and Martin et al. studies, on the other hand, used two different experimental groups, one for each test type. While the experimental groups were smaller in the McBride and Martin et al. studies than in the Kingsbury and Weiss study, McBride and Martin et al. used test items that had been calibrated with larger sample sizes and a better item calibration procedure. Moreover, the Martin et al. study was a replication of the McBride study. Martin et al. used the same conventional "predictor" test, the same Bayesian adaptive-test item pool, and the same 50-item conventional criterion test.

Since Martin et al. drew their experimental examinees from the same Marine recruit population that McBride sampled, it appears that any difference between the McBride and Martin et al. results are attributable to sampling error. While the similarity of the alternate-forms reliabilities and the adaptive-test validities obtained in the McBride and Martin et al. studies is encouraging, the observed difference (.87 versus .81) between the average conventional-test validities in these studies is rather large. In any event, simple averaging of the six conventional-test concurrent-validity coefficients represented in column three of Table 26 gives a value of .84 while averaging the six adaptive-test validity coefficients represented in column four gives a value of .83. This small average difference in favor of the six conventional tests is consistent with the results obtained for single-test validities in the present study.

The fourth study of adaptive-test validity (Thompson and Weiss, 1980) was unique in that it used nine different "real-world" criteria--high-school and college grade-point averages and ACT scores--instead of another word-knowledge test as the criterion. In the Thompson and Weiss study, two relatively small experimental groups (Groups 1 and 2) were tested. In Group 1, a variable-length "stradaptive" word-knowledge test was compared to a 40-item conventional word-knowledge test. Since there were two predictor tests and the stradaptive test was parameterized two ways and scored four ways, and since nine criterion scores were available, a total of 180 pairwise contrasts between individual validity coefficients were conducted in Group 1. The N for each comparison ranged from 55 to 101 since all nine criterion scores were not available for every examinee. A total of 21 statistically significant contrasts between validity coefficients were obtained. Four of these significant contrasts favored the adaptive test over the conventional test when overall college grade-point average was the criterion. One of the significant contrasts favored the conventional test over the adaptive test when college mathematics grade-point average was the criterion. The other 16 significant contrasts involved comparisons between various scorings and parameterizations of the adaptive test.

In Group 2 of the Thompson and Weiss study, a variable-length Bayesian word-knowledge test was compared to the same 40-item conventional word-knowledge test that was used in Group 1. In Group 2, the same nine criterion variables were used, but only one method of parameterizing and scoring the adaptive test was considered. Among the nine contrasts between adaptive- and conventional-test validities that were tested (with N varying from 71 to 131) only one was statistically significant. This contrast was in favor of the adaptive test when predicting high-school grade-point average.

Two considerations make interpretation of the results of the Thompson and Weiss study problematic. First, the significance testing procedure that was adopted in this study provided no control over the experimentwise error rate in either group of examinees. Since all contrasts within a group used overlapping sets of examinees, and since many of the contrasts differed only with regard to the adaptive-test scoring method or the item parameterization method involved, it may be concluded that most of the contrasts would be highly correlated over samples. Whenever contrasts are highly correlated and at least one Type I error is present in a sample, substantially more than 100α percent of the set of contrasts tested will be statistically significant at the α level, even if the null hypothesis is true for all of the contrasts.

In particular, the four significant contrasts in favor of the stradaptive testing strategy that were observed in Thompson and Weiss' Group 1 could be expected to either be all significant or all non-significant in any given replication of this study, since the adaptive-test predictor scores that were involved in these contrasts are known to correlate in the high nineties. (It should be emphasized that these comments do not serve to demonstrate that the four significant contrasts in favor of the stradaptive test were Type I errors. They do indicate that these four contrasts should be viewed as essentially one significant result, not four different significant results.)

The second problem associated with the Thompson and Weiss study is that a careful examination of the item parameters in the 40-item conventional test (their Appendix Table C) and the adaptive tests' item pools (their Appendix Tables A and B) reveals that the adaptive tests' pools contained a substantial number of items of distinctly higher quality than were present in the conventional test. In particular, if the 15 most discriminating items in Stratum 5 of the stradaptive-test item pool were combined with the 14 most discriminating items in Stratum 4, the result would be a 29-item conventional test that provides about 50% more test information near $\theta = 0$ than Thompson and Weiss' 40-item conventional test. Twenty-nine items closely approximates the mean stradaptive-test length that was observed in this study.

Similarly, a peaked 35-item conventional test with a mean discrimination parameter of approximately .80 (a value near the mean of the Bayesian-test item-pool discriminations) would provide about 90% more information near $\theta = 0$ than the 40-item conventional test (which had a mean discrimination value of approximately .54). Thirty-five items was the median Bayesian-adaptive-test length observed in this study.

Taken together, the results summarized in Table 26, the results of the Thompson and Weiss study, and the results of the current research seem consistent with the conclusion that there is as yet no clear-cut evidence that at a test length of 30 items either conventional or adaptive word-knowledge tests have consistently higher concurrent or criterion-related validities in the populations studied. Possibly the most important observation to make is that the adaptive tests did not have significantly lower validities than the conventional tests in these populations. This is not a trivial conclusion in light of the fact that adaptive tests administer different items to different examinees.

It is still possible that increases in predictive validity due to the use of adaptive testing strategies will be observed when these strategies are used in a full-range sample from the military applicant (AFEES) population. Research studies conducted to date, including the present one, have sampled examinees from populations with a narrower range of ability than is found in the AFEES population. Since the psychometric advantages of adaptive testing increase as the range of ability in the examinee population is increased, and since research in relatively narrow-range populations has demonstrated that adaptive testing does not degrade validity, the next logical step would be to compare the criterion-related validity of conventional and adaptive tests in a sample from the AFEES population.

Conclusions

Although this study did not demonstrate any statistically significant increases in criterion-related validity due to adaptive tests, it did support the feasibility of adaptive testing in military testing environments, since validities obtained using adaptive tests were not significantly different from those obtained from ASVAB subtests. In addition, the data showed that adaptive tests could provide levels of measurement precision obtainable only with much longer ASVAB tests, and that adaptive tests one-third to one-half the length of conventional ASVAB tests could approximate the criterion-related validities of these conventional tests. When combined with other advantages of computerized adaptive testing, including immediate availability of test results for selection and classification decisions, potential beneficial psychological effects (e. g., Betz & Weiss, 1976a, 1976b; Johnson et al., 1981), and the alleviation of test compromise problems, this study supports the potential utility of computerized adaptive ability testing in a military testing environment.

REFERENCES

Armed Forces Vocational Testing Group. Counselor's manual: Armed Services Vocational Aptitude Battery (Vol. 1). DOD 1304.12X. Washington, DC: U.S. Government Printing Office, 1974.

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1973.

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing. Research Report 75-3. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1975.

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance. Research Report 76-3. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (a)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing. Research Report 76-4. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program June 1976. (b)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Bock, R. D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1975.

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries. Research Report 77-6. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum, 1975.

Collier, R. O., & Hummel, T. J. Experimental design and interpretation. Berkeley, CA: McCutchan, 1977.

Draper, N. R. & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Dunn, O. J., & Clark, V. Correlation coefficients measured on the same individuals. Journal of the American Statistical Association, 1969, 64, 366-377.

Educational Testing Service. Cooperative School and College Ability Tests: A brief. Princeton, NJ: Educational Testing Service, Cooperative Test Division, 1958.

Educational Testing Service. Taking the SAT. Princeton, NJ: College Entrance Examination Board, 1978.

Fruchter, D. A., & Ree, M. J. Development of the Armed Services Vocational Aptitude Battery Forms 8, 9, and 10. AFHRL-TR-77-19, AD-A039270. Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, March 1977.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Jensen, H. E., Massey, I. H., & Valentine, L. D. Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7). AFHRL-TR-76-87, AD-A037522. Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1976.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1977, 1, 111-120.

Johnson, M. F. & Weiss, D. J. Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psycnology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Johnson, M. F., Weiss, D. J., & Prestwood, J. S. Effects of immediate feedback and pacing of item presentation on ability test performance and psychological reactions to testing. Research Report 81-2. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, February 1981.

Kendall, M. G., & Stuart, A. The advanced theory of statistics (Vol. 2, 3rd ed.). London: Griffin, 1973.

Kingsbury, G. G., & Weiss, D. J. An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests. Research Report 80-5. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, December 1980.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole, 1968.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1974.

Lord, F. M. Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-217. (a)

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. Research Bulletin 75-33. Princeton NJ: Educational Testing Service, 1975. (b)

Lord, F. M. Panel discussion: Future directions for computerized adaptive testing. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Martin, J. T., McBride, J. R., & Weiss, D. J. Reliability and validity of adaptive vs. conventional tests in a military recruit population. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory. (in press)

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J. R. Adaptive mental testing: The state of the art. Technical Report 423. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, November 1979.

McBride, J. R. Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.

Owen, R. J. *A Bayesian approach to tailored testing*. Research Bulletin 69-92. Princeton, NJ: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70*, 351-356.

Rao, C. R. *Linear statistical inference and its applications* (2nd ed.) New York. Wiley, 1973.

Ree, M. J. Estimating item characteristic curves. *Applied Psychological Measurement*, 1979, *3*, 371-385.

Sachar, J. D., & Fletcher, J. D. Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17, 1969.

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, 1977, *42*, 163-191.

Samejima, F. *Is Bayesian estimation proper for estimating the individual's ability?* Research Report 80-3. Knoxville: University of Tennessee, Department of Psychology, July 1980.

Swaminathan, H., & Gifford, J. Estimation of parameters in the 3-parameter latent trait model. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Sympson, J. B. Evaluating the results of computerized adaptive testing. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects*. Research Report 75-5. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

Sympson, J. B. *An alternating-least-squares procedure for fitting the logistic response model*. Paper presented at the Tenth Annual Mathematical Psychology Meeting, Los Angeles, August, 1977. (a)

Sympson, J. B. Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), *Applications of computerized testing*. Research Report 77-1. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1977. (b)

Sympson, J. B. *Testing differences between multiple correlations*. ETS RR-79-20. Princeton, NJ: Educational Testing Service, December 1979.

Sympson, J. B. *Estimating the reliability of adaptive tests from a single test administration*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Sympson, J. B. *Multivariate linear-model analysis of test validity coefficients*. Manuscript in preparation.

Thompson, J. G., & Weiss, D. J. *Criterion-related validity of adaptive testing strategies*. Research Report 80-3. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, June 1980.

Vale, C. D. Strategies of branching through an item pool. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects*. Research Report 75-5. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

Vale, C. D., & Weiss, D. J. *A study of computer-administered stradaptive ability testing*. Research Report 75-4. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975.

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. *Applied Psychological Measurement*, 1977, *1*, 141-152.

Waters, B. K. Discussion: Session 1. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980.

Weiss, D. J. & Betz, N. E. *Ability measurement: Conventional or adaptive?* Research Report 73-1. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Wood, R., Wingersky, M., & Lord, F. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

APPENDIX:

Supplementary Tables

Table A-1
Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive AR Tests in the Fixed-Entry Condition as a Function of Number of Items Administered, for AR-WK Group (N=221)

Test Length	BAYES Mean	BAYES SD	BAYES r	STMI* Mean	STMI* SD	STMI* r
0	-1.300	.000	.000	-1.300	.000	.000
1	-.751	1.048	.250	-.691	.781	.250
2	-.572	1.547	.347	-.435	1.037	.307
3	-.501	1.445	.371	-.307	1.300	.344
4	-.534	1.343	.400	-.256	1.302	.372
5	-.585	1.240	.431	-.258	1.264	.403
6	-.556	1.196	.422	-.239	1.208	.440
7	-.555	1.167	.457	-.222	1.144	.461
8	-.522	1.153	.474	-.227	1.127	.479
9	-.477	1.154	.446	-.239	1.115	.485
10	-.452	1.130	.448	-.237	1.103	.473
11	-.443	1.121	.469	-.233	1.078	.481
12	-.442	1.123	.457	-.248	1.084	.485
13	-.435	1.122	.457	-.249	1.084	.497
14	-.418	1.111	.457	-.268	1.124	.472
15	-.408	1.113	.455	-.264	1.132	.480
16	-.383	1.097	.457	-.268	1.114	.484
17	-.388	1.098	.459	-.274	1.100	.485
18	-.381	1.102	.455	-.278	1.086	.490
19	-.383	1.096	.467	-.276	1.065	.498
20**	-.382	1.093	.467	-.281	1.062	.501
21	-.384	1.094	.468	-.295	1.066	.497
22	-.382	1.084	.468	-.303	1.069	.499
23	-.378	1.079	.472	-.303	1.051	.501
24	-.388	1.071	.479	-.307	1.053	.493
25	-.386	1.075	.483	-.311	1.061	.489

*Original boundaries were used for maximum likelihood estimates.

**ASVAB/N validity at 20 items was .491.

Table A-2

Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive WK Tests in the Fixed-Entry Condition, as a Function of Number of Items Administered, for WK-AR Group (N=231)

Test	BAYES			STMI*		
Length	Mean	SD	r	Mean	SD	r
0	-1.400	.000	.000	-1.400	.000	.000
1	-.376	.875	.169	-.945	.436	.169
2	-.429	1.436	.100	-.537	.638	.231
3	-.298	1.171	.156	-.124	.993	.238
4	-.230	1.083	.161	-.009	1.169	.252
5	-.269	1.060	.191	-.033	1.084	.269
6	-.243	.989	.183	-.035	1.016	.282
7	-.200	.969	.186	-.041	.986	.290
8	-.201	.938	.224	-.068	.960	.290
9	-.185	.928	.238	-.088	.943	.288
10	-.206	.909	.237	-.092	.924	.280
11	-.205	.906	.244	-.117	.910	.273
12	-.199	.903	.248	-.113	.894	.273
13	-.200	.899	.259	-.112	.886	.272
14	-.197	.891	.259	-.131	.880	.265
15	-.206	.882	.272	-.145	.878	.254
16	-.194	.873	.268	-.143	.871	.262
17	-.188	.874	.259	-.143	.874	.266
18	-.195	.861	.262	-.139	.868	.268
19	-.189	.862	.272	-.147	.862	.273
20	-.188	.855	.274	-.147	.857	.268
21	-.186	.852	.272	-.149	.856	.269
22	-.180	.845	.282	-.153	.855	.275
23	-.183	.839	.280	-.155	.851	.271
24	-.180	.837	.287	-.156	.855	.276
25	-.175	.840	.290	-.158	.855	.276
26	-.176	.841	.291	-.161	.852	.276
27	-.181	.842	.289	-.164	.853	.272
28	-.178	.838	.289	-.161	.854	.271
29	-.177	.839	.295	-.160	.853	.275
30**	-.171	.838	.294	-.163	.850	.276
31	-.172	.841	.290	-.165	.855	.281
32	-.172	.842	.293	-.167	.858	.280
33	-.165	.840	.298	-.169	.854	.281
34	-.167	.840	.300	-.169	.854	.285
35	-.165	.836	.297	-.170	.854	.285

*Original boundaries were used for maximum likelihood estimates.

**ASVAB/N validity at 30 items was .294.

Table A-3

Means and Standard Deviations of Ability Estimates
and Criterion-Related Validity Correlations (r)
of BAYES and STMI Adaptive AR Tests in the
Variable-Entry Condition, as a Function of Number of
Items Administered, for WK-AR Group (N=231)

Test	BAYES			STMI*		
Length	Mean	SD	r	Mean	SD	r
0	-.620	.460	.297	-.685	.427	.285
1	-.448	1.108	.305	-.418	.807	.233
2	-.402	1.234	.336	-.365	1.067	.265
3	-.466	1.121	.337	-.383	1.335	.312
4	-.568	1.103	.327	-.355	1.277	.343
5	-.589	1.084	.364	-.354	1.196	.341
6	-.569	1.038	.361	-.346	1.145	.347
7	-.596	1.032	.353	-.374	1.162	.340
8	-.584	1.052	.361	-.388	1.129	.385
9	-.562	1.048	.392	-.408	1.141	.405
10	-.553	1.036	.409	-.393	1.097	.415
11	-.537	1.032	.403	-.397	1.080	.426
12	-.525	1.030	.407	-.398	1.069	.426
13	-.526	1.029	.408	-.386	1.016	.425
14	-.518	1.056	.404	-.420	1.077	.407
15	-.513	1.052	.417	-.431	1.110	.371
16	-.510	1.043	.432	-.438	1.115	.385
17	-.520	1.050	.443	-.442	1.115	.387
18	-.529	1.056	.450	-.448	1.107	.384
19	-.532	1.046	.451	-.455	1.105	.398
20**	-.535	1.044	.457	-.486	1.155	.406
21	-.538	1.044	.460	-.487	1.169	.378
22	-.536	1.045	.465	-.488	1.147	.396
23	-.539	1.038	.470	-.499	1.146	.402
24	-.541	1.037	.470	-.506	1.157	.417
25	-.542	1.036	.473	-.507	1.149	.425

*Original boundaries were used for maximum likelihood estimates.

**ASVAB/N validity at 20 items was .448.

Table A-4
Means and Standard Deviations of Ability Estimates and Criterion-Related Validity Correlations (r) of BAYES and STMI Adaptive WK Tests in the Variable-Entry Condition, as a Function of Number of Items Administered, for AR-WK Group (N=221)

Test Length	BAYES Mean	BAYES SD	BAYES r	STMI* Mean	STMI* SD	STMI* r
0	-.794	.709	.483	-.851	.584	.489
1	-.394	1.175	.257	-.453	.782	.413
2	-.263	1.288	.191	-.212	.936	.349
3	-.124	1.106	.224	.071	1.144	.314
4	-.093	1.063	.259	.047	1.143	.297
5	-.125	1.051	.255	.022	1.097	.316
6	-.165	1.016	.306	-.016	1.091	.313
7	-.175	.993	.297	-.036	1.028	.345
8	-.188	.978	.293	-.060	1.032	.333
9	-.183	.965	.301	-.049	.995	.344
10	-.181	.960	.299	-.048	.983	.330
11	-.167	.960	.306	-.058	.963	.333
12	-.153	.951	.305	-.063	.958	.333
13	-.145	.946	.319	-.066	.946	.333
14	-.153	.944	.322	-.083	.949	.346
15	-.158	.941	.314	-.090	.944	.344
16	-.152	.937	.323	-.097	.938	.347
17	-.153	.939	.338	-.097	.940	.340
18	-.145	.934	.340	-.091	.935	.343
19	-.145	.933	.340	-.095	.933	.347
20	-.138	.931	.343	-.091	.926	.345
21	-.132	.933	.346	-.095	.927	.347
22	-.132	.922	.344	-.099	.924	.347
23	-.127	.914	.345	-.102	.917	.343
24	-.129	.909	.342	-.108	.914	.344
25	-.132	.910	.342	-.105	.913	.341
26	-.135	.910	.345	-.105	.916	.338
27	-.133	.910	.346	-.110	.916	.339
28	-.134	.907	.344	-.109	.916	.342
29	-.135	.908	.345	-.111	.913	.343
30**	-.138	.904	.347	-.109	.911	.346
31	-.140	.901	.346	-.115	.906	.348
32	-.143	.898	.347	-.114	.908	.350
33	-.142	.894	.355	-.118	.906	.355
34	-.145	.898	.355	-.117	.904	.357
35	-.150	.900	.354	-.115	.904	.359

*Original boundaries were used for maximum likelihood estimates.

**ASVAB/N validity at 30 items was .409.

☆U.S. GOVERNMENT PRINTING OFFICE: 1982- 569-054/25